4-2018

Towards an Interactionist Dualism

Yonghao Wang

Towards an Interactionist Dualism

Yonghao Wang

College of William and Mary

Towards an Interactionist Dualism

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Arts in Philosophy from
The College of William and Mary

by

Yonghao Wang

Accepted for High Honors
(Honors, High Honors, Highest Honors)

Dr. Matthew Haug, Director

Dr. Joshua Gert

Dr. Robert Leventhal

Williamsburg, VA
April 25, 2018

Table of Contents

Introduction

The purpose of this honors thesis is to argue for interactionist dualism, the view that mental entities do not metaphysically supervene on physical entities and that they have causal efficacy in the physical world. This thesis is opposed to two theories. First, it denies epiphenomenalism, the idea that mental entities are effects but not causes. Second, and most importantly, it rejects physicalism. There is no consensus about the definition of physicalism, but most people agree that physicalism at least entails that mental entities metaphysically supervene on physical entities (the notion of supervenience is further discussed below). This honors thesis argues against this necessary condition of physicalism.

David Chalmers advocates for naturalistic dualism and is very positive toward Russellian panprotopsychism. Naturalistic dualism is a version of dualism that claims that the mental naturally, though not metaphysically, supervenes on the physical, and it does not commit to either epiphenomenalism or interactionist dualism. Russellian panprotopsychism is the view that there is a single kind of intrinsic, categorical feature of both physical and phenomenal properties, and this feature is called proto-phenomenal property for its similarity to phenomenal properties, though explanatorily speaking, proto-phenomenal properties are fundamental and give rise to phenomenal properties (as well as physical properties). Even though this theory is a form of monism, Chalmers argues that it is quite different from physicalism (Chalmers 1996:155). Essentially, Chalmers regards naturalistic dualism and panprotopsychism as compatible and tries to insert the theories as the middle ground among interactionist dualism, epiphenomenalism, and physicalism.

This thesis consists of two parts. In Part 1 I explain why we should endorse an interactionist version of dualism instead of other kinds of anti-physicalism. I argue that

epiphenomenalism is false, and as versions of anti-physicalism, naturalistic dualism and Russellian panprotopsychism fail to find the middle ground between interactionist dualism and epiphenomenalism. John Perry (2001) and Andrew Bailey (2006) have accused Chalmers of presupposing epiphenomenalism. I develop their attack based on the causal closure problem and reply to Chalmers' responses. In particular, I evaluate Chalmers' defense of Russellian panprotopsychism and argue that this theory as well cannot avoid committing to either interactionist dualism or epiphenomenalism. In Part 2 I provide an interactionist argument against physicalism by constructing a possible world which is only partially physically identical to the actual world. I then give an account of what mental causation in the interactionist dualist picture might look like and address some objections against interactionist dualism and the interactionist argument.

Part 1: Why Interactionist Dualism?

The Logical Space of the Causal Closure Problem

The causal closure problem can be roughly formulated as follows, "Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?" (Kim 1998:38) On the surface, we need to reconcile two seemingly contradicting claims. On one hand, we have strong evidence for what is called the completeness of physics, the thesis that every physical effect is fully caused by previous physical occurrences. On the other hand, it is very intuitive that our consciousness, or our mental activities, plays a causal role in our physical life. If we further take two other views, that physical effects are not overdetermined by both mental and physical causes, and that mental properties are not physical and do not supervene on the physical, we have four logically incompatible claims:

a. Every physical effect is fully caused by previous physical occurrences. (the completeness of physics)

b. Mental properties have physical effects. (the negation of epiphenomenalism)

c. Physical effects are not overdetermined by both mental and physical causes. (no overdetermination)

d. Mental properties are not physical and do not supervene on physical properties. (dualism)

Again, (c) claims that physical effects only have one sufficient cause at a given time. (b) claims that physical effects can have mental causes, while (a) claims that physical effects always have physical causes. (d) states that mental properties are different from physical properties, and therefore mental causes are not the same as physical causes. These four claims are clearly incompatible. If every physical effect is guaranteed a physical cause, and it is allowed only one cause, then there is no room for mental causes which are not, or do not supervene on, physical causes. Many physicalists find it most sensible to reject (d), claiming that mental properties are ultimately physical or supervene on the physical (but of course, a physicalist does not have to accept (b) or (c)).

Epiphenomenalism

To be a dualist means to accept (d), and a property dualist has to deny one of the first three propositions. Among the options, to deny (b) is to accept epiphenomenalism, which denies that mental properties have causal efficacy in the physical world. While this is certainly not impossible, many philosophers reject it as counterintuitive. For example, John Perry offers an example of him eating a delicious cookie, having the wonderful phenomenal experience of the

taste and exclaiming "Boy, was that good!" Perry claims that "I find it simply incredible—not inconceivable, but really quite incredible—that the conscious event was not part of the cause of my saying what I did" (Perry 2001:74).

Andrew Bailey (2006) offers three arguments against epiphenomenalism, and the one I find most convincing is the reporting problem, that epiphenomenalism seems incompatible with the self-reporting of consciousness. "If consciousness is epiphenomenal then it has no effects; in particular, it has no effects on those organisms whose consciousness it is" (Bailey 2006:494). If this is true, then my experience of pain does not cause, and is completely causally irrelevant to my utterance of "my leg hurts," which seems highly unlikely.

It is noteworthy that Chalmers has responded to an epistemological problem of epiphenomenalism, which is similar to, but essentially different from the reporting problem. According to Chalmers, the most influential form of this problem is raised by Sydney Shoemaker (1975). The problem states that if mental states are not physical and do not have causal efficacy, then they do not cause our beliefs about them; but if our beliefs about our mental states are not caused by our mental properties, we are not justified in having those beliefs; therefore, either mental states have causal efficacy, or we are not justified in having beliefs about our mental states. In response to this challenge, Chalmers denies the causal theory of knowledge, at least as it applies to our knowledge of our mental states. The causal theory of knowledge claims that the justification for believing *p* requires a causal relation between the fact that *p* and the belief that *p*. Chalmers contends that this causal relation is not a necessary condition. Rather, justification in the case of mental beliefs requires only that we be directly acquainted with the mental states.

I do not go into details about this response because, even if this is true, it only accounts for the justification of beliefs about mental states but fails to explain what causes the beliefs in

the first place. To make sure we are talking about a physical effect here, the final effect is the physical utterance of the sentence "my leg hurts," which is presumably caused by my belief about my mental state of my leg hurting. Even granted that the justification for having the belief is not causal, this physical effect still needs to be caused. Whereas the utterance is, at least partially, caused by the belief, the belief seems to be caused by the mental state. Therefore, the reporting problem seems to pose a real problem for epiphenomenalism, and to deal with the causal closure, rejecting (b) would not be a reasonable move for dualist if there are better options.

Overdetermination

For dualists, an apparent way to avoid epiphenomenalism is to deny (c) (no overdetermination) instead. If physical effects are overdetermined by mental properties and physical properties, then even if epiphenomenalism is false, the causal efficacy of mental properties and the completeness of physics can then coexist. It is noteworthy that (c) is a strong statement that claims there are no cases of overdetermination at all. Generally, people reject systematic overdetermination, since this seems like a coincidence that has too small a probability to be considered seriously. However, it is not clear we have any reason to deny all cases of overdetermination. After all, overdetermination is perfectly possible and some random cases of it do not seem to face the same problem as systematic overdetermination does. Therefore, if we can find just one random case of overdetermination, then it seems we have successfully disproved (c).

This seems a possible solution to the causal closure problem, but upon closer inspection, simply introducing some random cases of overdetermination does not help. Suppose a dualist wants to solve the problem with this method without admitting epiphenomenalism or rejecting completeness. Since completeness is still true, one has to maintain that all physical effects of

causally efficacious mental properties have sufficient physical causes as well. If this is true, then mental properties would be "causally irrelevant in that their presence or absence makes no difference to how everything (non-phenomenal) goes in the world," which Bailey labels as "quepiphenomenalism" (Bailey 2006:488). However, "the problems [with quepiphenomenalism] arise because of the implausibility of supposing that the elimination of consciousness would make no difference at all to how things go in the physical world" (ibid.). This argument adopts the same, though a stronger, intuition as Bailey and Perry did when refuting epiphenomenalism, stronger in the sense that it claims that mental properties have to make a difference to the physical world, instead of just having causal efficacy. To use overdetermination, random or systematic, as a way out is to claim that removing any mental properties has no physical effects at all, which is as unlikely. Therefore, random or systematic overdetermination alone cannot solve the causal closure problem. As a dualist who rejects epiphenomenalism, one has to deny the completeness of physics.

The Zombie Argument

Chalmers is known for his zombie argument for dualism. Philosophical zombies are supposedly possible beings which are physically identical to humans, but which have no mental experience. Chalmers develops the zombie argument in his (1996) book *The Conscious Mind*. The argument goes roughly as follows:

1. Zombies are conceivable.
2. Whatever is conceivable is metaphysically possible.
3. Therefore zombies are metaphysically possible.
4. If zombies are metaphysically possible, then dualism is true.

5. Therefore, dualism is true.

Again, physicalism entails that mental entities metaphysically supervene on physical entities. When I use "supervenience" in this essay I mean strong supervenience, for which Jaegwon Kim gives the following definition: "let A and B be families of properties…, A strongly supervenes on B just in case, necessarily, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and necessarily if any y has G, it has F" (1984:164-5). Simply speaking, physicalism is true in a world only if, necessarily, whenever a mental property, M, is instantiated, there is also a physical property, P, that is instantiated in the same entity, and necessarily, whenever P is instantiated, M is also instantiated in the same entity. The zombie argument, however, tries to show that it is possible that the same physical properties as in the actual world are instantiated but mental properties are not, from which follows that in the actual world, mental properties do not supervene on physical properties. If this is true, then mental properties are fundamental, and property dualism is true.

Perry and Bailey accuse the zombie argument, rightfully in my view, of denying (b) and accepting epiphenomenalism. A zombie world is supposed to be physically identical to the actual world, and therefore zombies are possible only if removing mental properties from humans has no physical consequences. This already sounds a lot like epiphenomenalism. On a simple inspection, it is difficult to deny epiphenomenalism while accepting the possibility of zombie worlds. As Perry asks, "[i]f conscious states make a difference in the way our bodies work and ultimately in how we behave, and they are absent in the zombie world, then how could everything in the physical world be the same as it is in our world?" (2001:73). Similarly, Bailey argues that if zombie world is possible, then "on the face of it, consciousness is not required for

everything to happen in the actual world just as it does, and so consciousness is radically epiphenomenal (in the actual world)" (Bailey 2006:488).

I have been talking above about why epiphenomenalism is a bad choice for dualists, making an interactionist dualism very desirable. However, this also means that dualists have to deny the completeness of physics, a thesis that is widely accepted. Chalmers seems convinced by the evidence that supports the completeness of physics (1996:125), and in order to avoid rejecting completeness or committing to epiphenomenalism, Chalmers tries to expand the logical space of the causal closure problem.

Chalmers' Solution 1

In his review of John Perry's book, Chalmers offers two possible ways to avoid epiphenomenalism in the zombie argument without rejecting the completeness of physics:

> … [A]n interactionist dualist can accept the possibility of zombies, by accepting the possibility of physically identical worlds in which physical causal gaps go unfilled, or are filled by something other than mental processes. The first possibility would have many unexplained physical events, but there is nothing metaphysically impossible about unexplained physical events. Also: a Russellian "panprotopsychist", who holds that consciousness is constituted by the unknown intrinsic categorical bases of microphysical dispositions, can accept the possibility of zombies by accepting the possibility of worlds in which the microphysical dispositions have a different categorical basis, or none at all. (Chalmers 2004:184)

The first method states that we might be able to stipulate the zombie world to be physically identical to the actual world. Suppose (just for the sake of argument and brevity) that pain causes me to yell in the actual world. In the zombie world, we can stipulate that I yell at the same time and in the same manner as I do in the actual world, only that my yelling is not caused by anything, or caused by something other than mental or physical properties (maybe an alien type of property). With this solution, it seems that we can construct a possible zombie world that

is physically identical to the actual world, but on a closer look, as Chalmers changes the cause of the physical effects (me yelling), the physical laws in the zombie world would be different from those in the actual world. In the same paper as above, Bailey elaborates on Chalmers' defense:

> Suppose we elect not to fill the causal gaps but simply stipulate that the physical events continue to occur as they do in the target world. This cannot be done without changing the physics. As Chalmers himself is often at pains to point out, the characterization of the physical is structural and relational; what makes an electron an electron, as far as the physical sciences are concerned, is the way it is embedded in a set of law-like causal relationships with other entities…. [I]t is impossible for two physical events, one connected by natural laws to other event-types and the other not so connected and hence 'unexplained,' to be the same physical event (i.e. members of the same physical event-type). (Bailey 2006:493)

Here Bailey adopts a specific view, that structural and relational properties are both essential for the identity of the physical. An electron is not an electron if it does not have the same causal relationship with other charged entities. The causal efficacy of an electron determines its identity. This view is also supported by Chalmers himself. If this is true, then it seems that the zombie world is not physically identical to the actual world if one chooses to try to stipulate that the physical facts in the zombie world are the same as those in the actual world, while some physical events that are caused by physical events in actual world are uncaused (or caused by alien properties) in the zombie world. In the actual world some physical events stand in causal relationships with the mental, whereas in the zombie world they stand in causal relationships with something else or even nothing. The zombie argument only works when it proves the possibility to have the same physical properties without having the same mental properties. Bailey's argument, if sound, shows that Chalmers fails to construct a zombie world where the same physical properties as in the actual world are instantiated.

Chalmers' Solution 2: Russellian Panprotopsychism

The other method suggested by Chalmers is to be a Russellian "panprotopsychist." On this view, "consciousness is closely tied to the intrinsic properties that serve as the categorical bases of microphysical dispositions" (Chalmers 2010:151). One of the motivations for this idea is that if physical properties are just relational properties, then the world seems "strangely insubstantial" (Chalmers 1996:153). Therefore, it seems reasonable to suppose that there are intrinsic properties of the physical as the "categorical basis," and the relational or causal properties relate these intrinsic bases. The only intrinsic properties we really know of are phenomenal properties. Since it is better to give a simple, unified explanation of both the physical and the mental, a way to achieve this is to posit a proto-phenomenal feature that is the categorical feature of both physical and mental properties. As Chalmers notes, "[i]t is natural to speculate that there may be some relation or even overlap between the uncharacterized intrinsic properties of physical entities, and the familiar intrinsic properties of experience" (1996:154) On this view, proto-phenomenal features "give rise to" or "aggregate to" phenomenal properties that we know of, and combined with causal, relational features, they together give rise to or aggregate to physical properties.

With this theory, Chalmers tries to make room for the causal efficacy of mental while remaining committed to the completeness of physics:

> If there are intrinsic properties of the physical, it is instantiations of these properties that physical causation ultimately relates. If these are phenomenal properties, then there is phenomenal causation; and if these are protophenomenal properties, then phenomenal properties inherit causal relevance by their supervenient status, just as billiard balls inherit causal relevance from molecules…
>
> …the intrinsic properties should not be identified with physical properties such as mass. It seems reasonable to say that there is still mass in the zombie world,

> despite differences in its intrinsic nature. If so, then mass is an extrinsic property that can be 'realized' by different intrinsic properties in different worlds. (ibid.)

The causal efficacy of mental properties is presented as follows: events of physical causation causally relate two instantiations of the categorical bases of the physical, i.e. proto-phenomenal properties. Because one of the relata of causation is the cause, proto-phenomenal properties are then causally efficacious in that they can cause the instantiations of other proto-phenomenal properties. Since phenomenal, or mental, properties derive from proto-phenomenal properties, they inherit the causal efficacy of the proto-phenomenal properties.

The next step is to apply this picture to the zombie argument. What Chalmers needs is to establish that the physical world can remain the same in the zombie world. On the surface, when you remove all phenomenal properties, what you essentially remove is their categorical bases, i.e. proto-phenomenal properties, but when the proto-phenomenal properties are removed, then there seems to be no categorical basis for physical properties. Chalmers argues that some alien type of intrinsic properties, or no properties at all, could be the intrinsic bases of physical properties, and physical properties are identified simply by their relational, extrinsic properties. If this is right, then even though he changes the categorical bases of physical properties, he still manages to keep everything physical the same between the actual world and the zombie world while keeping the causal efficacy of mental properties in the actual world.

There are at least two places where we can raise doubt about this account. First, it is unclear that Chalmers can justifiably say that physical properties are identified solely by their relational, extrinsic properties. Chalmers argues that "it seems reasonable to say that there is still mass in the zombie world" when the intrinsic features of mass are changed, but this is like saying there is still water in Putnam's Twin earth. If a thing consisting of XYZ is not water even if it behaves exactly like water (Putnam 1973), how can we say a thing consisting of a different

intrinsic feature than mass is still mass even though it behaves exactly like mass? If this is right, then Chalmers commits a similar mistake here as in his first solution by failing to keep the actual world and the zombie world physically identical. While with the first method he changes the relational features of the physical, with the second one he changes the intrinsic features of the physical.

Of course, water is a kind of substance and mass is a kind of property, so there is a category difference in these two cases. One might then argue that even though at non-fundamental categories, the composition of an entity is essential for its identity, it might be the case that at the fundamental level, what is essential is only the relational features, and at the intrinsic features are not essential for the identity of fundamental physical entities. For example, the identity of quarks, if it is not composed of any more fundamental entities, might depend only on its dispositional, relational properties. While this is certainly possible, it seems implausible, for if they are not necessary for the identity of fundamental physical entities, then why introduce them in the first place? Since Chalmers allows that there be no categorical bases for physical entities without changing the identity of those physical entities, then there is no reason to suppose that they do have categorical bases in the actual world. The world would indeed seem "insubstantial," but there is nothing wrong with being insubstantial, for we do not observe substantiality anyway, and being substantial is almost trivial for being physical. If this is right, then this move to maintain the identity of physical entities after changing its categorical bases undermines the motivation to introduce categorical bases of physical entities.

Now, even if intrinsic features are not essential to the physical, Chalmers still faces a more serious problem, which is that the causal efficacy that phenomenal properties gain on this account does not really enable one to deny epiphenomenalism while maintaining completeness.

Chalmers claims that phenomenal properties inherit causal efficacy from proto-phenomenal properties, but what is this inherited causal efficacy like? If it is full-blown efficacy that makes a physical difference, then completeness is wrong, for some physical effects would only have mental causes. If it is not full-blown efficacy, then it cannot account for our intuition against epiphenomenalism that our mental experience does play a causal role. The intuition we have in the reporting problem is that normal phenomenal properties seem to cause physical events and make a physical difference, which necessarily contradicts completeness if one accepts dualism. Therefore, there is no satisfactory way to expand the logical space of the causal closure problem. Since we reject epiphenomenalism, one has to adopt an interactionist account for dualism to be true.

Part 2: Why dualism in the first place?

Now that we have seen that it is more reasonable to believe in interactionist dualism than non-interactionist dualism (epiphenomenalism or Russellian panprotopsychism), the task remains to show that it is more reasonable to believe in dualism than physicalism. In this section I will offer what I call the interactionist argument for dualism. This argument is a modified version of the zombie argument. The zombie argument tries to argue that no mental properties supervene on physical properties by constructing a possible world in which no mental properties are instantiated. However, the possibility of the zombie world entails that removing mental properties has no physical effect in the actual world, and therefore the zombie world is actually not possible because epiphenomenalism is false in the actual world. Though the zombie argument does not disprove physicalism, we need to note that physicalism is a rather fragile

claim, as any mental property that does not supervene on physical properties can prove it wrong. Therefore, I argue against physicalism by constructing a possible world in which only one mental property is absent.

The Interactionist Argument

The interactionist argument tries to have only one mental property not instantiated in a possible world (I select pain arbitrarily), and if this argument shows that this one mental property does not supervene on physical properties, then physicalism is false.

Since it is argued above that epiphenomenalism is false, then what is going on in our mind plays a causal role, and, for example, the pain that I feel causes me to shout, "Ouch!" Let's then conceive of a world A, where physical and mental facts remain as similar to the actual world as logic allows. World A and the actual world have the same physical properties and substances (e.g. mass, charge, quarks, stones), and they have the same physical laws. Plants grow and water (H_2O) flows as in the actual world. The essential difference between the two worlds is that whereas pain (call it M) is instantiated in the actual world, it is not in world A. Other mental properties are still instantiated in world A, so some animals (say prehistoric fish or birds, though we are not sure whether it is earlier, simpler organisms that first experience pain; humans would likely not evolve due to their ancestors' lack of reaction to pain) still have phenomenal experiences, but no conscious beings feel pain at all. World A then differs physically from the actual world at least in that it does not have the physical events that pain brings about. For example, the sensation of pain when coming really close to fire in the actual world causes birds to avoid coming so close to fire in the future, but in world A, presumably, they do not avoid doing so even after coming close to fire for the first time. Of course, the physical difference that

the absence of pain causes will probably cause other physical differences, making world A much different from the actual world, but that does not concern us. Again, a physicalist has to maintain that necessarily, whenever we have a mental property M (say pain) instantiated in an entity (e.g. a bird), we have a physical property P (say some sort of neuron firing) instantiated there, and necessarily, whenever P is instantiated in an entity, M is instantiated there as well. The aim of this argument is to show that it is possible that P is instantiated while M is not.

Now consider the first instantiation of pain (call it m_1) in the actual world, which happens at t_1. m_1 is not present in world A, and the key work is to establish that the first instantiation of P (call it p_1) is present in world A. As said earlier, we are keeping world A as similar to the actual world as logic allows. What logic does not allow is that while pain is not instantiated in world A, the physical effects of pain are present in world A, but whatever is not the effect of pain (or of the effects of pain) can be the same in both worlds. It is clear that p_1 is not the effect of m_1 or of the effect of m_1, since there is no reason to suppose causes (like p_1) ever supervene on effects (like m_1). Therefore, p_1 can be safely stipulated to be present in world A. With the same reasoning, we can conclude p_1 is not the effect of the effects of m1.

If this is right, then it is perfectly conceivable that p_1 is instantiated in a world where m_1 is not instantiated, and if conceivability entails possibility, M does not supervene on P and property dualism is true in the actual world. The overall argument is formalized as follows:

1. It is conceivable to have p_1 (some particular sort of neuron firing) without m_1 (the first instantiation of pain).
2. Whatever is conceivable is possible.
3. It is possible to have p_1 without m_1.
4. Therefore, pain does not supervene on neuron firing.

5. Therefore, property dualism is true.

This new argument acknowledges the causal efficacy of mental properties and therefore avoids epiphenomenalism. At the same time, this argument does not change the identity of p_1 in world A, since neither its causal relation nor its intrinsic feature needs to be altered in world A in order for world A to be possible. It therefore avoids the same kind of problem the zombie argument faces, that, as Bailey complains, by stipulating the physical events to be the same, Chalmers changes the identity of the physical events already.

Dualist Accounts of Mental Causation

With an argument for interactionist dualism in hand, now we need to explain what mental causation is like. After all, our current understanding of causation is mostly about physical causation, and it is important to provide a coherent, plausible account of causation and to show how mental causation is different from physical causation, if different at all.

One cannot avoid discussing energy when discussing causation. Most, if not all, cases of physical causation seem to involve energy. A frequently used example of physical causation is when one billiard ball hits another one; in this case, energy from the first billiard ball transfers to the second one. An important aspect of energy is the conservation of energy (henceforth CoE), which states that in a closed system, i.e. a system which does not exchange energy with other systems, the total amount of energy always remains the same. CoE is a well-established physical law, and physicalists question how mental causation is possible, given that our universe is a closed system in which energy is conserved. Now, CoE itself is logically compatible with interactionist dualism, so attackers need to add additional premises to reject dualism. Ben White

suggests that one such premise could be "any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered" (2017:389), but he also points out that this is not quite enough, for if one allows mental causes to possess energy, the contradiction would vanish. Therefore, White further suggests that one more premise be added, either that "nothing non-physical has energy (or at least none that is capable of being transferred to any physical body) or … that the physical realm constitutes a closed system (i.e., one that exchanges no matter or energy with its surroundings, and on which no external force acts)" (ibid.). The first option rules out any mental energy at all, or at least energy that would interact with the physical; the second one prevents the physical from interacting with anything non-physical, including anything mental.

Essentially, both statements deny that the mental can interact with the physical, at least not by means of energy. Therefore, dualists need to show either that the mental and the physical can indeed interact by means of energy, or that they can interact by some other means. To make the task clearer, dualists are dealing with this set of incompatible claims:

e. Our universe is a closed system that conserves energy.
f. Any change in a body's motion involves some transference of energy between the cause of the change and the body whose movement is altered.
g. Nothing non-physical has energy (or at least none that is capable of being transferred to any physical body).
h. Mental states can cause a body's motions.

Among these claims, (e) is well supported by scientific evidence, but (f) and (g) are not. To save (h), dualists can therefore either reject (f) and claim that the mental can cause a body's motion without itself possessing energy or transferring energy to the body, or reject (g) and claim

that the mental can possess energy and thus interact with the physical by means of energy. I will discuss the second option first, as this is the approach White takes. I believe White's approach is a possible explanation of mental causation, but it is implausible and the first option is the better way for dualists to take.

Can the Mental Possess Energy?

The starting point of White's approach is that "it is unclear how something could exert a force without possessing energy" (2017:391). He does not explicitly explain the reason he believes so, but in a later section, he claims "energy is partly defined in terms of force (for energy is the capacity to do work or transfer heat, and work is the application of force to a body that results in the displacement of that body in the force's direction)" (392). The idea seems to be that the concepts of energy, work, and force are closely related, and these concepts are essential to our understanding of how causation works. Many theories of causation would support his claim as well. For example, a version of the conserved quantity theory claims that "a causal interaction is an intersection of world lines which involves exchange of a conserved quantity" (Dowe 1995:323). On this view, if the mental has a causal interaction with the physical, the mental must transfer a conserved quantity to the physical, namely a conserved amount of energy. In order to defend this position, White needs to make sense of mental energy.

A general worry is that energy is a property attributed to physical entities, and we have no account of energy possessed by non-physical entities. White argues that "there is no reason why a non-physical entity could not be ascribed a physical quantity if such an ascription were warranted by certain effects that it was found to have upon some physical system" (2017:391). The idea is that if dualism is true and after a mental event, the total amount of energy possessed

by physical entities is found increased, then it seems more reasonable to suppose that there is energy possessed by non-physical entities as well, instead of rejecting CoE immediately. We do not have a definitive reason to suppose that only physical entities can possess energy, and the transferability of this quantity possessed by non-physical entities to energy possessed by physical entities suggests that the quantity is energy as well.

While this account does address some worries, it nonetheless has a serious problem. First, White does not clarify whether "energy possessed by physical entities" and "energy possessed by non-physical entities" are numerically identical to each other. It seems that this should be the case, for they are unified by the same CoE. Just like how chemical energy and kinetic energy are both energy, only in different forms, "energy possessed by physical entities" and "energy possessed by non-physical entities" should also both be energy, only possessed by two different kinds of entities.

However, White grants in a footnote that "on the assumption that while non-physical entities might possess energy, they cannot possess mass, the energy possessed by non-physical entities would have to differ from that possessed by physical entities at least in the respect that when possessed by non-physical entities, it is not equivalent with mass" (ibid.). Now this is puzzling. Energy is equivalent to mass whether in the form of chemical energy or in the form of kinetic energy, but this explanation suggests that while energy is equivalent to mass when possessed by physical entities, it is not equivalent to mass when possessed by non-physical entities. How could this be? To say that energy possessed by non-physical entities and that possessed by physical entities are the same energy would violate the indiscernibility of identicals, roughly the principle that *a* and *b* are numerically identical only if they have exactly the same properties. If this is right, then it seems energy possessed by non-physical entities and that

possessed by physical entities cannot be numerically identical, and White is positing a new category of beings, namely mental energy, instead of simply energy possessed by non-physical entities. This new kind of energy would be fundamentally different from physical energy that is equivalent to mass, but somehow mental energy could transfer to physical entities and transform to physical energy that is equivalent to mass, and mental energy and physical energy are regulated by the same CoE despite their fundamental difference. This seems highly implausible, if not completely impossible. White's note therefore at least contradicts himself, and in order to make sense of mental causation, dualists have to adopt another approach.

Can the Mental Have Physical Effects Without Itself Possessing Energy?

Since White's account is unlikely to be true, how exactly can the mental have physical effect, if not by means of energy? Let's take the example of me speaking the word "Pizza" when asked what my favorite food is. This is a physical event/effect that transfers the energy stored in my body into kinetic and thermal energy. Between the event someone asking the question and me speaking the word, there are also some intermediate causes. To describe the whole causal chain in more detail, and from the perspective of an interactionist dualist, the man's voice when asking the question causes certain vibration of my eardrums; some brain states about the sound input taken then happen; my mental representation of the information is formed; I then have an intention to speak; then some brain state that instructs the vocal muscles to move happens; and finally electrical signals are sent to the muscles, my mouth moves, and my vocal folds vibrate.

Now the fine-grained part of the mental causation here is the conscious, mental state of intention causes some brain state that instructs the vocal muscles to move. More specifically, the brain state is an electric signal forming and travelling from the central nervous system to the

peripheral nervous system. This electric signal requires energy to form and travel, but does the energy have to come from the mental cause? The answer is negative, for the chemical energy stored in the body is already sufficient. This is different from cases of physical causation, e.g. a billiard ball hitting another one. In this case of physical causation, the cause provides the energy required for the effect, but in the case of mental causation, the body provides the energy required for the effect. The mental cause only functions to trigger the transformation from chemical energy to electrical and kinetic energy, and it is not clear that this triggering necessarily requires energy. If this is true, then energy transference from the cause to the effect is not required for all cases of causation. As White mentions, C. D. Broad first claims that mental causes "determine that at a given moment so much energy shall change from the chemical form to the form of bodily movement; and they determine this, so far as we can see, without altering the total amount of energy in the physical world" (1925:109). This way, we have a coherent account of mental causation without attributing energy to mental entities.

It is noteworthy that I do not argue that all cases of apparent mental causation are actually cases of mental causation. For all we know according to the best of science, many of our actions are not at all caused by any conscious activity. Maybe my description of the above case is wrong. Maybe no mental intention is needed for me to spontaneously speak the word, which might not be caused by anything mental. But my thesis is rather that at least some cases of apparent mental causation involve intermediate mental causes, and without these mental causes, the physical effects would not happen. It does seem that at least in cases of self-reporting, the involvement of conscious activity as the cause is required. Can I really say "I believe it is 8 o'clock in the morning" when my belief that it is 8 o'clock in the morning is not the cause? The same intuition

we have about epiphenomenalism applies here. If this is right, then we have good reason to believe that there are some cases where mental states have physical effects.

The biggest challenge for this account is that we do not know how this causal relation works. I claim that mental properties can trigger physical energy transformation and transference without energy, but I have not provided an account of how this "triggering" works. In a case of physical causation, e.g. a signal from the central nervous system causing a body movement, we explain the causal relation by means of energy, claiming that it is the energy carried by the signal which transfers to the muscles that triggers the body movement. In my example of mental state causing a signal to form and travel, however, we have no explanation of what triggers this signal to form and travel. Process theorists, e.g. conserved quantity theorists, would therefore contend that this account is not plausible because there is no exchange of a conserved quantity, i.e. energy, between the cause and the effect that serves the triggering role.

However, this objection can stem simply from our ignorance of the nature of causal processes in general. This response presupposes that there is only one kind of causal process, that which physicalists understand it to be. But imagine a possible world where only two states of affairs obtain, namely a mental (non-physical) state and a physical brain state. The mental state does not possess any energy, but by some weird natural law in that world (and there is indeed a law), and through some weird process different than energy transference, that mental state "influences" and "gives rise to" the occurrence of that physical state. Now, the question for process theorists is as follows: given that there is a process (though a weird one), by what right could one deny that this is a causal relation? Surely, we do not know yet about what the causal process is here, but it seems that the only thing the process theorists can say is that such

causation is not the physical causation we know of. Our ignorance of possible causal processes cannot warrant a rejection of unknown causal processes.

Think about mental causes and mental effects. My beliefs that if *P* then *Q* and that *P* cause my belief that *Q*. What is the process behind this case of causation? What triggers the formation of my belief that *Q*? We really have no idea, and we cannot explain this mental-mental causation just by saying that beliefs are mental and of course they can have causal effect on each other, just as we cannot explain physical causation just by saying they are physical. Still, it is not reasonable to say that therefore, a mental belief cannot cause another belief. Rather, we say that a mental belief can cause another even though we do not know about the causal process behind this causal relation. In that case, since it is reasonable to suppose that there is a causal process that we do not know of in the case of mental-mental causation, it is also reasonable to suppose so in the case of mental-physical causation as well. Therefore, this account of mental causation without appealing to energy is still plausible.

The Place of Mental State in Nature

Chalmers holds naturalistic dualism, roughly the idea that mental states naturally, or nomologically, supervene on physical brain states. The problem with this account is that if mental states were to supervene on physical brain states, then it seems mental states are causally efficacious not in virtue of themselves, but in virtue of the physical brain states. If you remove the mental states, nothing physical would change, which is contradictory to our intuition about quepiphenomenalism, the idea that mental states do not make a physical difference (whether they are causally efficacious or not). Therefore, if the interactionist account is right, then it is not the case that mental states somehow emerge from brain states according to natural laws, or the case

that whenever you have a mental state, you also have a brain state at the same time. Mental states are caused by brain states, and they cause brain states as well. They just do not naturally supervene on brain states.

Is this already disproved by science, or can science tell us that there is always some brain state when you have a mental state? I think not. While I do think we have strong evidence for the correlation between mental and brain states, we might not have any evidence to suppose that this correlation is supervenience or causation. To acquire such evidence, you need to determine whether the mental states and the brain states are simultaneous. If we discover that they are indeed simultaneous, then we would have strong evidence for the supervenience claim. However, the only way you can know when a mental state happens is by self-reporting, but self-reporting of time can never be exact. Humans have a reaction time, and however close the reported time when the mental state happens and the recorded time when the brain state happens are, we cannot be sure whether they happen at the same time or not. After all, the time between the cause and the effect might just be a few milliseconds, probably far shorter than human reaction time. Therefore, it is difficult to see how we can ever determine whether the mental supervenes on the physical by determining the time which correlation between mental and brain states is. Therefore, to reject interactionist dualism, physicalists need to find other weakness in the argument and account.

Objections to the Interactionist Argument and Interactionist Dualism

When Perry and Bailey attack the zombie argument, they say that zombies are inconceivable because it entails epiphenomenalism in the actual world. The interactionist argument does not face this problem, but physicalists, especially identity theorists, can still argue

that world A is not conceivable either because in the actual world, mental properties are identical to physical properties, and therefore when you remove pain, you remove something physical already. Some also deny that conceivability entails possibility using Saul Kripke's idea of a posteriori necessity. As Chalmers himself has mentioned, "it is often said that sentences such as 'water is not H2O' provide counterexamples to the claim that conceivability entails possibility: it is conceivable that water is not H2O, but it is not metaphysically possible" (2010:145). Therefore, apparently, we have an example of something conceivable but impossible. One thing to notice here, though, is the claim that water is necessarily H2O is an identity claim. Chalmers generally addresses the second objection with the two-dimensional argument, which is beyond the scope of this thesis. I will instead focus on the first objection based on identity theories.

To respond to this objection, I want to turn to the motivation for identity theories and point out that these theories are brought up because of and justified by a belief in physicalism. According to Kim, there are three positive arguments for physicalism in general or identity theories in specific. I want to raise doubt for all three arguments for physicalism and thereby diminish the threat to dualism and the interactionist argument by identity theories.

The first two argue for identity theories directly. "The first, originally promoted by Smart without much elaboration, is the simplicity argument, to the effect that identifying mental states, including states of consciousness, with neural/physical states of the brain, helps us attain the simplest, most parsimonious worldview" (Kim 2005:124). However, as Kim correctly points out, "[w]hat a physicalist may seize upon as the most parsimonious and elegant ontology would be apt to strike the dualist as a hopelessly inadequate scheme which discards, or ignores, the entities that are needed to save the phenomena" (ibid.:125). This simplicity argument is therefore not able to support strong theses like the identity theories on its own.

The other argument is what Kim calls the explanatory argument. Kim specifies two kinds of the explanatory argument in his (2005) book. The first one is advanced by Christopher Hill and Brian McLaughlin. The central idea is that, based on inference to the best explanation, the best explanation for the correlation between mental and brain states is that they are identical, and therefore an identity theory is probably true (or we have good reason to believe that it is true). Against this argument, Kim brings up at least four points. First, Kim suggests that "[e]ven if we were to grant that type physicalism is to be preferred over its rivals, the warrant it enjoys might be far from sufficient for it to merit our 'outright' belief or acceptance" (ibid.:128). An identity theory might indeed best explain the correlation between mental and brain states, but it is not clear that its explanatory power is so strong that it is sufficient for us to believe the theory. Sometimes, when we compare two theories explaining the same phenomena, one might only be slightly better than the other, and this slight advantage might not be able to grant a belief in that theory.

Second, when we apply the principle of inference to the best explanation, the best theory should not only best explain the correlation between mental and brain states, but also other issues in philosophy of mind. For example, dualism seems to best explain the possibility of the zombie world or that of the qualia inversion. When comparing the explanatory power over all relevant issues, identity theories do not have a clear advantage over dualism. Third, the data that theories explain are changing constantly. Future data can influence how strong the explanatory power of a theory is, and therefore new data could undermine identity theories' explanatory power. Fourth, Kim argues that a reduction to identity is not really an explanation of correlation. It is simply a reiteration of existing phenomena. He uses the example of Tully, who is identical to Cicero. If Tully is wise, then we can safely claim that Cicero is wise, but this is not an explanation of why

Tully is wise. It only restates the fact that Tully is wise in a different way. Therefore, Kim believes this version of explanatory argument cannot support the identity theories.

Ned Block and Robert Stalnaker adopt a quite different version of the explanatory argument, claiming that "acceptance of [some] identities is sufficiently justified because they enable explanations that mere correlations could not yield" (ibid. 141). For example, we have a good physiological explanation of how a particular brain state causes another brain state. If we identify pain with the former state and stress with the latter, then we have a good explanation of how pain can cause stress. This account looks nice at first, but Kim points out that in this picture, the work of explanation is done by neuropsychology, and the only thing identity theories do is to redescribe in folk vocabulary a phenomenon that has already been explained [in neuropsychology]" (ibid. 146). The above example seemingly explains how pain causes stress, but in fact it only states in a different way (the colloquial way) the fact that a certain brain state causes another brain state. Therefore, nothing new is explained by identity theories, and without any explanatory power, identity theories are not well supported.

Based on the objections above, Kim believes the causal argument is the way for physicalists to go. The causal argument argues for physicalism in general and is basically the causal closure problem I have described in Part 1: given that completeness is true, that epiphenomenalism is false, and that systematic overdetermination is false, physicalism must be true. Therefore, the strongest support for identity theory lies in completeness, the very thesis dualists have to deny in the causal closure problem. The remaining task is then to argue that even completeness is not as strong as physicalists believe it to be, and therefore that there is no adequate positive argument for physicalism or identity theories.

In his essay "The Rise of Physicalism," David Papineau explains two popular arguments for the completeness of physics based on empirical findings:

> (1) The Argument from Fundamental Forces. The first argument is that all apparently special forces characteristically reduce to a small stock of basic physical forces that conserve energy. Causes of macroscopic accelerations standardly turn out to be composed out of a few fundamental physical forces that operate throughout nature. So, while we ordinarily attribute certain physical effects to 'muscular forces,' say, or indeed to 'mental causes,' we should recognize that these causes, just as all causes of physical effects, are ultimately composed of the few basic physical forces.
>
> (2) The Argument from Physiology. The second argument is simply that there is no direct evidence for vital or mental forces. Physiological research reveals no phenomena in living bodies that manifest such forces. All organic processes in living bodies seem to be fully accounted for by normal physical forces. (2001:27)

Both arguments, in my opinion, are not as strong as people think they are. The first argument claims that upon examining the history of science, all apparently special forces reduce to a few basic physical forces that conserve energy, so it is reasonable to suppose that mental forces would reduce to some fundamental physical forces as well. However, we need to note that "force" is not synonymous to "cause." As I have discussed in a previous section, mental states can be causally efficacious without having any energy or force. This argument, therefore, at best shows that all forces can be reduced to fundamental physical forces. It does not show that all causes are ultimately physical causes.

The second argument claims that science (especially physiology) has yet to discover any mental or non-physical forces, so there is no reason to suppose there are any. Besides my last point that mental states can be causally efficacious without having energy or force, we can also argue that even if we have not found mental causes in any scientific research, this still does not provide enough evidence against mental causes. As White accurately points out, "our scientific understanding of the inner workings of the brain is currently still in its initial stages" (2017:397).

Physiology and neuroscience are still very young disciplines, and it is still quite possible that there be new discovery of mental causes in the future. Maybe in the future we would discover that without a certain mental state, some physical state would not happen. Even though the current evidence is probably enough to justify the inclination to believe that all forces are physical, it is not enough to argue against the possibility of undiscovered mental causes, as Papineau attempts to do. Therefore, this argument also fails to support the completeness of physics.

I have shown above that the three positive arguments are not all strong as physicalists believe they are, and therefore we have less reason to think identity theories are true. Since there is no definitive rejection of identity theories either, now we have a somewhat mutually question-begging situation. Interactionist dualists claim that mental states are not identical to physical brain states, and identity theorists claim otherwise. Which side you support largely depends on how strong you believe in the three arguments vs. how strong you believe in your intuition that the mental is not physical. It seems to me that neither can give a knock-down argument against the other. However, given how prevailing physicalism is in contemporary philosophy, this is already a progress for interactionist dualism. Furthermore, since dualism can accommodate both the scientific evidence that is supposed to support completeness, and our intuition about the mind, dualism seems to have a stronger explanatory power.

Conclusion

The two starting points of interactionist dualism are: 1) mental states are NOT physical and 2) mental states have causal efficacy in the physical world, that is, their presence makes a

physical difference. When one only believes the first proposition, he can be a physicalist and say minimally that mental states supervene on physical brain states. When one only believes the second claim, he can be an identity theorist and say mental states just are physical brain states. However, physicalists cannot believe both propositions, for given that mental states supervene on physical brain states, it seems the causal efficacy belongs to the physical brain states, and removing mental states does not seem to make a difference to the physical world. This honors thesis also respects scientific discoveries about the mind, and I have argued above that scientific discoveries do not provide sufficient reasons to support the completeness of physics or to reject either of the above propositions, and that it is more reasonable to believe in interactionist dualism.

References

Bailey, Andrew R. (2006). Zombies, epiphenomenalism, and physicalist theories of consciousness. *Canadian Journal of Philosophy* 36 (4):481-509.

Broad, C. D. (1925) *The Mind and Its Place in Nature*. New York: Harcourt, Brace & Company, Inc.

Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, David J. (2002). Does conceivability entail possibility? In Tamar S. Gendler & John Hawthorne (eds.), *Conceivability and Possibility* (pp. 145--200). Oxford: Oxford University Press.

Chalmers, David J. (2004). Imagination, indexicality, and intensions. *Philosophy and Phenomenological Research* 68 (January):182-90.

Chalmers, David J. (2010). The 2-Dimensional Argument against Materialism. In *The Character of Consciousness* (pp. 141-205). Oxford: Oxford University Press.

Chalmers, David J. (2013). Panpsychism and Panprotopsychism. *The Amherst Lecture in Philosophy* 8 (2013): 1–35. <http://www.amherstlecture.org/chalmers2013/>.

Dowe, P. (1995). Causality and Conserved Quantities: A Reply to Salmon. *Philosophy of Science* 62: 321-333.

Kim, Jaegwon (1984). Concepts of supervenience. *Philosophy and Phenomenological Research* 45 (December):153-76.

Kim, Jaegwon (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

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

Papineau, David (2001). The rise of physicalism. In Carl Gillett & Barry M. Loewer (eds.), *Physicalism and its Discontents* (pp. 3-36). Cambridge: Cambridge University Press.

Perry, John (2001). *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.

Putnam, Hilary (1973). Meaning and reference. *Journal of Philosoph*y 70 (19):699-711.

White, Ben (2017). Conservation Laws and Interactionist Dualism. *Philosophical Quarterly* 67 (267):387–405.