Introduction

> I am not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is. Whereas, if it is not literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[1]
>
> Jerry A. Fodor (1989: 77)

0.1 The Problems of Mental Causation

Mental causation is causation by mental causes. More specifically, it is the causation of physical effects by mental causes. In this book, I will use 'mental causation' in this specific sense. Nothing as fanciful as causing a spoon to bend through sheer force of will is required for mental causation. Rather, there is mental causation whenever what is going on in our minds causes our bodies to move. This, it seems, happens all the time. I have a headache and take an aspirin. In a typical case, my headache causes my hand to move towards the aspirin. If an argument for the causal claim is needed, here is one: in a typical case, my hand moves towards the aspirin because I have a headache. Whatever 'because' means precisely in this context, it at least implies that my headache causes my hand to move. More generally, it seems that there is mental causation whenever we act intentionally. Whatever it means precisely to act intentionally, it at least implies that some of our mental states cause our bodies to move.[2]

[1] Jerry Fodor, excerpt from 'Making Mind Matter More', *Philosophical Topics* 17, no. 1 (1989). Copyright © 1989 by the Board of Trustees of the University of Arkansas. Reprinted with the permission of The Permissions Company, LLC on behalf of the University of Arkansas Press.

[2] If there are purely mental intentional actions, at least the causation of mental effects by mental causes is implied.

Thus, both common sense and philosophical theorizing have it that what is going on in our minds at least sometimes causes things to happen in the physical world. All the same, there are philosophical problems about mental causation. This book deals with the two most serious ones: the interaction problem and the exclusion problem. The interaction problem is the problem of how there can be mental causation at all. How, in principle, can the mental causally interact with the physical? If we think of the causing of physical effects as a job, is the mental qualified to do this job?[3] The exclusion problem is the problem of how there can be mental causation given that all physical effects already have physical causes. How could physical effects have additional mental causes? Even if the mental is qualified to cause physical effects, how can it do this job if the physical is already doing all the work? Perhaps there could be some kind of job-sharing between mental and physical causes, such that some physical effects have physical and mental causes that somehow act in tandem. But then, it seems, the situation would be similar to a firing squad[4] where the victim's death is overdetermined by the firings of two shooters, and it seems implausible that physical effects are thus overdetermined whenever there is mental causation. The interaction problem and the exclusion problem are connected. The exclusion problem arises only if the interaction problem has been solved. The mental can compete with the physical for the job of causing mental events or share this job with the physical only if the mental is qualified to do the job.

How serious the interaction and exclusion problems are depends on the nature of mind. The more intimate the relation between mind and body, the less pressing both problems become. The most intimate relation is that of identity. Reductive physicalists claim that the mental is identical to the physical. If reductive physicalism is true, then the interaction problem disappears. If the mental is identical to the physical, then the mental causes of physical effects are *ipso facto* physical causes, and no one doubts that physical effects can have physical causes.[5] The exclusion problem disappears too. If the mental is identical to the physical, there are no additional mental causes, since any mental causes simply *are* physical causes. Physical effects of mental causes are not caused twice over, leading to cases of

[3] My presentation of the relation between the interaction and exclusion problems and my use of the job metaphor loosely follows Karen Bennett's (2007: 325, 2008: 281 n. 4).

[4] I shall follow the established tradition of using the somewhat macabre example of the firing squad. In the philosophy of causation, this tradition goes back at least to Mackie (1965).

[5] Almost no one: if one doubts that there is causation at all, as Russell (1912) does, *a fortiori* one doubts that physical effects can have physical causes.

overdetermination that resemble firing squads; rather, they have a single cause that is both mental and physical.

Non-reductive physicalists claim that the mental is distinct from the physical, while it is still metaphysically dependent on the physical. Metaphysical dependence can in turn be spelled out in different ways; most commonly, it is read as supervenience or asymmetric metaphysical necessitation. The relation of distinctness-cum-metaphysical-dependence is less intimate than the relation of identity. This opens a gap between the mental and the physical through which the interaction problem might enter. It does not, however, seem particularly problematic to see how physical effects could in principle have mental causes given that the mental metaphysically depends on the physical. The exclusion problem, on the other hand, looks serious. If the mental is distinct from the physical, any mental causes of physical effects that already have physical causes are additional causes of those effects. How can there be such additional mental causes without the physical effects' being overdetermined?

Dualists claim that the mental is distinct from the physical and does not even metaphysically depend on the physical. This relation is sufficiently loose to make the interaction problem pressing; it is also sufficiently loose to make the exclusion problem pressing if the interaction problem can somehow be solved. If the mental does not even metaphysically depend on the physical, how do the mental and physical realms interact causally at all? Even if they can interact in principle, how can mental and physical causes of physical effects coexist without yielding cases of overdetermination? The looser the connection between the mental and the physical, it seems, the more the case of a physical effect that has both a mental and a physical cause looks like a firing squad case.

While the severity of the interaction and exclusion problems depends on the nature of mind, it also depends on the nature of causation. The problems are severe if causation requires the transfer of some conserved quantity, such as energy, from the cause to the effect. If this is the case, then, in order to solve the interaction problem, dualists have to explain how something that does not metaphysically depend on the physical can transmit something to something physical. In order to solve the exclusion problem, dualists and non-reductive physicalists have to explain how both a physical cause and a distinct mental cause can transmit something to the same physical effect. These tasks seem hard, if not impossible.

By contrast, if it suffices for causation that one event makes a difference to another – in the sense that the second event would not have occurred had the first event not occurred – then the interaction problem and the

exclusion problem are manageable. In that case, in order to solve the interaction problem, dualists have to explain how the mental can make a difference to the physical. In order to solve the exclusion problem, dualists and non-reductive physicalists have to explain how both a physical cause and a distinct mental cause can make a difference to a physical effect without yielding a case of overdetermination or, at any rate, without yielding a case of overdetermination that is objectionable.

This book pursues the strategy of solving the interaction and exclusion problems through a difference-making account of causation. It argues that, given a difference-making approach to causation, dualists as well as non-reductive physicalists can solve the problems of mental causation. We do not even need a full-blown theory of causation that states necessary and sufficient conditions for causation in terms of difference-making in order to solve the problems for dualism and non-reductive physicalism. It suffices to assume that difference-making is sufficient for causation. This assumption is very plausible. Since it does not amount to a full-blown theory of causation, it is also relatively metaphysically weak. As we shall see, the minimal assumptions about the nature of mind that need to be made in order to solve the problems of mental causation through the strategy advocated here are also relatively weak metaphysically. They are compatible with a dualist position about the mind provided some fine-tuning is made about how easily mental phenomena and their physical bases could have come apart. Thus, the strategy pursued in this book uses weak assumptions both about the metaphysics of causation and about the metaphysics of mind. The metaphysical slack is picked up by the logic of the statements that express difference-making claims. These statements, counterfactual conditionals, obey distinctive logical principles that allow us to derive claims about how the mental makes a difference to the physical from claims about how the physical makes a difference to the physical, together with claims about the nature of mind. The derived claims about how the mental makes a difference to the physical can in turn be combined with the assumption about causation to derive claims about how mental events can have physical effects. In the debate about mental causation, the power of the logic of counterfactual conditionals has been overlooked so far. Instead, theories about the nature of mind have borne the weight of solving the problems of mental causation. Once the power of this logic is utilized, however, weak assumptions about causation as well as about the nature of mind suffice, and even dualists can account for mental causation.

The plan of this book is as follows. Chapter 1 lays the groundwork about the mind and about causation. It characterizes the different theories about

the nature of mind in more detail. It then turns to causation, its relata, and counterfactual conditionals. Counterfactual conditionals, their general truth-conditions and logical relations are introduced, as are more specific issues about how to evaluate them. The chapter defends a principle about causation in terms of counterfactual conditionals that will be crucial for later arguments. According to this principle, an event causes another (roughly) if the second event would not have occurred had the first event not occurred. While it is very plausible, the principle needs some refinement in order to deal with a number of *prima facie* difficulties. In particular, certain assumptions need to be made about how to evaluate counterfactual conditionals like 'If the first event had not occurred, then the second event would not have occurred.'

Chapter 2 uses the principle about causation to show that there are physical effects of mental causes. If non-reductive physicalism is true, applying the principle is straightforward. Indeed, it might seem that applying the principle is almost too straightforward, for the principle also yields non-mental higher-level causes that might be considered problematic. These causes are best diagnosed as causes that have little explanatory relevance. If dualism is true, applying the principle about causation in order to show the existence of mental causation is less straightforward, but still possible. In order to avail themselves of the principle, dualists need to assume that the laws that connect the mental and physical realms have a special status. In sum, both non-reductive physicalists and dualists can solve the interaction problem.

Chapter 3 shows that more sophisticated difference-making theories of causation that draw on so-called causal models can accommodate mental causation too. Causal modelling theories invoke more complex relations of difference-making than the simple principle about causation does. These relations are represented by causal models. Accommodating mental causation – either in the non-reductive physicalist case or in the dualist case – calls for some heterodoxy in model-building. If the heterodox models are allowed, however, they prove useful not merely for explaining mental causation, but also for capturing the distinction between higher-level causes that are explanatorily relevant and higher-level causes that are not.

Chapter 4 deals with the exclusion problem. If a difference-making approach to causation is adopted, the exclusion problem can be solved. Although mental and physical causes might nominally overdetermine their physical effects, these cases are sufficiently dissimilar to standard cases of overdetermination not to be problematic. Unlike in cases of mental

causation (on the account offered here), in standard cases of overdetermination, the individual causes do *not* make a difference to the effect.

The resulting picture of mental causation, summarized in Chapter 5, has repercussions for debates about the nature of mind. If virtually all theories about the nature of mind can solve the problems of mental causation, then arguments from mental causation against certain theories become irrelevant.

0.2 A Brief History of the Problems

The interaction problem took centre stage in the context of Descartes' view of the nature of mind.[6] Descartes held that human minds are souls and that souls and bodies are radically different – and hence distinct – substances, the soul being thinking and not spatially extended, the body being spatially extended and not thinking. Princess Elisabeth of Bohemia criticized Descartes for being unable to account for how the soul can initiate bodily movements. In a letter to Descartes, she wrote:

> So I ask you please to tell me how the soul of a human being (it being only a thinking substance) can determine the bodily spirits, in order to bring about voluntary actions. For it seems that all determination of movement happens through the impulsion of the thing moved, by the manner in which it is pushed by that which moves it, or else by the particular qualities and shape of the surface of the latter. Physical contact is required for the first two conditions, extension for the third. You entirely exclude the one [extension] from the notion you have of the soul, and the other [physical contact] appears to me incompatible with an immaterial thing.[7]

Descartes replied that we had a primitive notion of the union of the soul with the body, 'on which depends that of the power the soul has to move the body and the body to act on the soul, in causing its sensations and passions'.[8] Since the notion was primitive, it defied further explanation and could be 'understood only through itself'.[9] Elisabeth was not convinced. In her reply, she advanced what might be one of the earliest arguments from mental causation for the claim that the mind is material: 'I admit that it would be easier for me to concede matter and extension to the soul than to

[6] Debates about mental causation in antiquity are discussed in Caston 1997.
[7] Elisabeth to Descartes, 6 May 1643, AT 3:661, Princess Elisabeth and Descartes 2007: 62 (parentheses original). (Citations of the form 'AT x:y' refer to p. y of vol. x of Descartes 1996.)
[8] Descartes to Elisabeth, 21 May 1643, AT 3:665, Princess Elisabeth and Descartes 2007: 65.
[9] Descartes to Elisabeth, 21 May 1643, AT 3:666, Princess Elisabeth and Descartes 2007: 65.

concede the capacity to move a body and to be moved by it to an immaterial thing.'[10]

While he failed to attenuate Elisabeth's doubts, elsewhere Descartes elaborated on the locus of the interaction between soul and body.[11] Vital spirits swirled around the pineal gland, Descartes held, and could move limbs in a quasi-hydraulic manner. By moving the pineal gland, the soul could deflect the motions of the vital spirits and thus initiate bodily movements.[12]

Leibniz read Descartes as saying that the soul never changes the speed of anything, but causes bodily movements by changing the *direction* of the vital spirits. Thus, the soul can influence movements in the body 'much as a rider, though giving no force to the horse he mounts, nevertheless controls it by guiding that force in any direction he pleases'.[13] This account, Leibniz held, was consistent with Descartes' physics, which required merely that the 'quantity of motion' be conserved, where a body's quantity of motion is, roughly, the product of its mass and speed.[14] Leibniz held that his own physics ruled out that the soul could change the direction of a body's motion. He required not the conservation of quantity of motion, but the conservation of what is nowadays called momentum, where a body's momentum is the product of its mass and its velocity, velocity being a vector quantity. The soul's changing a body's direction of motion in the way Descartes – at least Leibniz's Descartes – envisaged would require changing the body's velocity vector and hence would violate the conservation of momentum.[15]

In addition to the conservation of momentum, Leibniz endorsed what is nowadays called the conservation of kinetic energy, where a body's kinetic energy is the product of its mass and the square of its speed. The conservation of momentum rules out that the soul changes the direction of motion

[10] Elisabeth to Descartes, 10 June 1643, AT 3:685, Princess Elisabeth and Descartes 2007: 68.

[11] For further discussion of the correspondence between Elisabeth and Descartes, see Garber 1983b and Perler 1996: 123–160.

[12] See Descartes, *Passions of the Soul*, §§34, 41; AT 11:354–11:355, 11:359–11:360; CSM 1:341, 1:343. (Citations of the form 'CSM x:y' refer to p. y of vol. x of Descartes 1984–1991.)

[13] *Theodicy*, §60 (Leibniz 1985: 156).

[14] In fact, Descartes defined a body's quantity of motion as the product of the body's speed and its *size*, which in turn is closely related to, but not identical with, its volume; see Slowik 2014, §4. On the face of it, Leibniz's reading of Descartes seems charitable, but Descartes may not have intended his physical laws to hold without exception. For instance, in the *Principles of Philosophy* he explicitly restricts his third law of motion to causes of corporeal changes that are themselves corporeal and brackets human and angelic minds as possible causes (see AT 8:65, CSM 1:242). See Garber 1983a for further discussion.

[15] See *First Explanation of the New System*, §20 (Leibniz 1898: 327–328).

of a single particle. In principle, it leaves open that the soul interacts with bodies in a different way. For instance, when bodies collide, the conservation of momentum by itself leaves open different post-collision situations, depending on whether the collision is elastic or inelastic. (In the inelastic case, it leaves open different post-collision situations in turn, depending on how inelastic the collision is.) In elastic collisions, kinetic energy is conserved; in inelastic collisions, it is not. One might hypothesize that the soul changes the motion of bodies by causing certain collisions to be elastic and others to be inelastic. If we stipulate that kinetic energy as well as momentum be conserved, that option is ruled out. Indeed, Leibniz thought that once both the conservation of energy and the conservation of momentum are in place, the soul cannot interact with bodies at all: 'without a complete derangement of the laws of Nature the soul could not act physically upon the body'.[16] In contemporary terminology, Leibniz accuses Descartes of falling prey to the exclusion problem: given the nature of the physical world, there is no room left for anything non-physical to bring about physical changes.

Some modern commentators (e.g., Papineau (2001: 15)) have concurred with Leibniz's claim that the two conservation laws fully fix the future position and velocity of bodies and thus rule out any non-physical influence on the motion of bodies. Strictly speaking, this claim is not true, however. Take two point-size particles of equal mass and speed that are about to collide head-on. We would expect their post-collision velocity vectors to be the negatives of the pre-collision velocity vectors. But even if we assume that kinetic energy and momentum are conserved, any rotation of the expected post-collision situation is likewise possible (see Figure 0.1).[17]

Thus, Leibniz's conservation laws leave a loophole that Cartesians could try to exploit. This is not to say that the loophole cannot be closed by supposing further physical principles to hold, although finding suitable principles turns out to be surprisingly tricky.[18] (In order not to raise false

[16] *Theodicy*, §61 (Leibniz 1985: 156).

[17] In the example from the figure, the conserved momentum is $m\vec{v}_1 + m\vec{v}_2 = \vec{0}$, and the conserved kinetic energy is $\frac{1}{2}m|\vec{v}_1|^2 + \frac{1}{2}m|\vec{v}_2|^2 = m|\vec{v}_1|^2 = m|\vec{v}_2|^2$.

[18] See Gibb 2010 for discussion. Gibb holds that the classical conservation laws leave it open that momentum and energy be 'redistributed'. It is not clear whether the case illustrated in Figure 0.1 qualifies as a case of redistribution of momentum in her sense, for it is unclear whether there is any more 'redistribution' going on between $\vec{v}_1/\vec{v}_2$ and $\vec{v}'_1/\vec{v}'_2$ than between $\vec{v}_1/\vec{v}_2$ and $\vec{v}''_1/\vec{v}''_2$. For further discussion of the role of energy conservation in arguments against interactive dualism, see Averill and Keating 1981, Montero 2006, and Koksvik 2007. Presumably, Leibniz would have responded to the collision example that the $\vec{v}'_1/\vec{v}'_2$ and $\vec{v}''_1/\vec{v}''_2$ situations are identical because space is not absolute. This response could be blocked by adding another particle that is not involved in the collision. In the new scenario, an alternative response would still have been available to

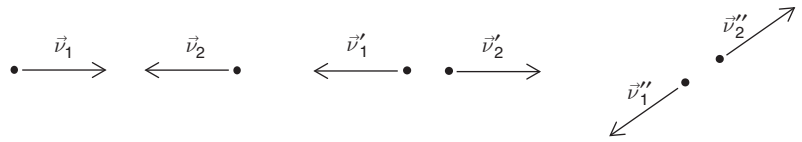

Figure 0.1. A collision ($\vec{v}_1/\vec{v}_2$) with different possible outcomes ($\vec{v}'_1/\vec{v}'_2$ vs $\vec{v}''_1/\vec{v}''_2$) despite conservation of momentum and kinetic energy

expectations, I should say that the account of mental causation I am going to recommend to dualists does *not* exploit any such loopholes.)

One way of sidestepping Elisabethian worries about the interaction between soul and body is to deny that this interaction is direct. One can hold that, instead, God acts as an intermediary whenever there seems to be a direct interaction between soul and body; in particular, God perceives our acts of will and brings it about that our bodies move accordingly. God is not subject to whatever limits there are on the direct interaction between soul and body and brings about their indirect interaction. This is the doctrine of occasionalism, which was held by Malebranche and other followers of Descartes.[19]

Leibniz rejected occasionalism just as he rejected Descartes' interactionism. His reasons for rejecting the two positions overlapped. Leibniz held that, similarly to Descartes' interactionism, the indirect interaction that occasionalism posits would violate the laws of nature. Moreover, occasionalism denigrated God to something like a theatrical *deus ex machina* who has to constantly interfere with his own creation.[20]

Leibniz's own answer to the question of how mind and body interact was that they did not interact at all, directly or indirectly. With interactionism and occasionalism ruled out, he wrote,

> there remains only my hypothesis, that is to say, *the way of the harmony pre-established* by a contrivance of the Divine foresight, which has from the beginning formed each of these substances in so perfect, so regular and accurate a manner that by merely following its own laws which were given to it when it came into being, each substance is yet in harmony with the other,

Leibniz. He could have claimed that there is no sufficient reason for the particles to assume the $\vec{v}''_1/\vec{v}''_2$ velocities, while there is a sufficient reason for them to assume the $\vec{v}'_1/\vec{v}'_2$ velocities, namely that the latter are parallel to the initial velocities.

[19] See Malebranche, Dialogue VII from *Dialogues on Metaphysics and on Religion* (Malebranche 1997: 104–126). Nadler (1997) argues that, contrary to the received view, Cartesians did not adopt occasionalism primarily in response to problems about mind–body interaction.

[20] See *Theodicy*, §61 (Leibniz 1985: 156). For further discussion of Leibniz's criticism of occasionalism, see Sleigh 1990 and Rutherford 1993.

> just as if there were a mutual influence between them, or as if God were continually putting His hand upon them, in addition to His general support.[21]

According to Leibniz, mind and body are like two clocks that are in perfect agreement with each other because they are made 'with such skill and accuracy that we can be sure that they will always afterwards keep time together'.[22]

Both occasionalism and the doctrine of pre-established harmony assign an indispensable role to God, as a constant causal mediator and meticulous creator, respectively. Once appealing to God in order to solve philosophical problems started to be considered problematic, new approaches to the interaction of soul and body were needed.

One approach was to deny that soul and body are distinct. Julien Offray de La Mettrie endorsed such a materialist view.[23] In his 1748 book *Man a Machine*, he defended the view that 'man is but an animal, or a collection of springs which wind each other up' (La Mettrie 1912: 135). In this machinery,

> the soul is but a principle of motion or a material and sensible part of the brain, which can be regarded, without fear of error, as the mainspring of the whole machine, having a visible influence on all the parts. (La Mettrie 1912: 135)

If the soul is nothing but a part of the physical machinery of our body, there is no problem of explaining how the soul can cause bodily movements.

Another approach was to accept that mind and body are distinct and that there is a causal relation between them, albeit a causal relation that goes merely one way, from body to mind. This is the doctrine of epiphenomenalism, whose chief proponent was Thomas Huxley (1874).[24] Huxley's epiphenomenalism incorporates an aspect of La Mettrie's view, viz. that the body is a complex machine that does not allow, or at least does not require, any non-physical influence. Indeed, Huxley's argument for epiphenomenalism rests mainly on empirical examples of complex bodily goings-on in the absence of consciousness. While there is a causal relation between body and mind according to epiphenomenalism, it is not in the direction from mind to body, which has generally been considered the

[21] *Third Explanation of the New System*, Leibniz 1898: 333–334. [22] Leibniz 1898: 333.

[23] Hobbes was an early materialist too, but, unlike La Mettrie, he was also a theist. See Springborg 2012 and Gorham 2013 for discussion.

[24] Other early proponents of epiphenomenalism include Clifford (1874) and Spalding (1877).

more troublesome direction. Thus, unlike occasionalism, epiphenomenalism need not posit any divine assistance for the causal relation.

Epiphenomenalism never became popular, either in the nineteenth or in the twentieth century; the materialist approach (which today is more commonly called 'physicalism') attracted more followers. Talk of the soul fell out of fashion, but identity theorists like Ullin Place (1956) and Herbert Feigl (1958) instead identified (types of) mental states with (types of) brain states. In the terminology of the previous section, these theorists qualified as reductive physicalists, although they did not apply this label to themselves. Like earlier materialist theories, the identity theory faces no problems from mental causation, for there are no problems with brain states causing bodily effects.

The success of the identity theory did not last long, however. Hilary Putnam (1967) pointed out that different physical states can underlie one and the same mental state, which rules out an identity between the mental state and any of those physical states. With the growing acceptance of this multiple realizability of mental states, which led to non-reductive physicalist theories of the mind, problems of mental causation re-emerged. Jaegwon Kim (1989), building on an argument by Norman Malcolm (1968), challenged advocates of non-reductive physicalism with a version of the exclusion problem. Physical effects already have physical causes. If they have additional mental causes – even mental causes that are tied to the physical causes by a relation of metaphysical dependence – then they are overdetermined; but a widespread overdetermination of the (putative) physical effects of mental causes was unacceptable, Kim held.

Besides arguments from multiple realizability, the identity theory faced an influential modal objection that was advanced by Saul Kripke (1980). Psychophysical identities had to be necessary if they obtained at all, Kripke held, but it seemed possible to be in the mental state without being in the (allegedly) identical physical state, and vice versa. It turned out that similar modal arguments could be advanced against non-reductive physicalism too (see Block 1978, Chalmers 1996). This led to a revival of dualism, although contemporary dualists, unlike their early modern predecessors, tend to hold that mental states or properties are distinct from physical states or properties, without going so far as to say that there are Cartesian souls.[25] Thus, the interaction problem is not

[25] According to a survey by Bourget and Chalmers (2014: 476–477), 27.1 per cent of professional philosophers believe in non-physicalism about the mind, and 23.3 per cent believe in the metaphysical possibility of zombies (that is, beings like us physically but without consciousness), which is standardly taken to entail dualism.

quite as severe for contemporary dualists as it was for Descartes, but it still seems hard to explain how mental states and properties, being neither identical to nor metaphysically dependent on physical states and properties, can have physical effects in principle. And even if this problem can be solved, Kim's exclusion problem stands in the way of a complete dualist account of mental causation.