Mental Causation

Holly Andersen, Simon Fraser University

Abstract

The problem of mental causation in contemporary philosophy of mind concerns the possibility of holding two different views that are in apparent tension. The first is physicalism, the view that there is nothing more to the world than the physical. The second is that the mental has genuine causal efficacy in a way that does not reduce to pure physical particle-bumping. This article provides a historical background to this question, with focus on Davidson's anomalous monism and Kim's causal exclusion problem. Responses to causal exclusion are categorized in terms of six different argumentative strategies. In conclusion, caution is advised regarding the inclination to reduce the mental to the physical and sketch a positive direction for substantively characterizing mental causation by recourse to well-confirmed accounts of causation coupled with empirical research.

Key words

causation, mind, physicalism, reduction, exclusion, anomalous monism

Introduction

Cognitive neuroscience is often take to either imply, or minimally to be compatible with, a view about the nature of mind called physicalism. Physicalism is the view that everything in the world, including but not limited to the mind, is purely physical in character; it essentially denies the existence of non-physical entities or processes. The debate about the existence and nature of mental causation stems from a tension between a commitment to physicalism concerning the nature of the mind, and the apparent causal efficacy of such mental events or states such as having an intention or committing an action. The commitment to physicalism is inconsistent with 'spooky' mental causes, such that thoughts, beliefs, or intentions are not themselves physical, but somehow reach out and poke the physical world. Physicalism is generally taken to imply that, whatever mindedness turns out to be, it should fit clearly into the nexus of physical causes with which we are already familiar. On one hand, this commitment appears very much in line with the approach to studying the mind in cognitive neuroscience. On the other hand, the realm of the mental

appears at least prima facie to be genuinely causally efficacious – we do at least *seem* to bring about our actions by deliberation and intentions on which we act – in ways that cannot be straightforwardly reduced to familiar physical causation. This perceived conflict between physical causation and the place of the mental is the source of debate about the existence and nature of mental causation.

The fact that the mental is characteristically ordered by rational norms that do not appear to exist in physical causation pulls towards causal autonomy of the mental from the physical, while the commitment to physicalism resists this as spooky or mysterious. Solving the problem of mental causation requires finding a way to reconcile physicalism about the mind with causal efficacy of the mental, either by situating mental causes in the physical world, by rejecting mental causes, or by rejecting physicalism. This article will focus on the first two of these options, since the commitment to physicalism is both widespread in this discussion, and a unifying premise shared by many disparate views.

The problem of mental causation is closely related to, but not the same as, what is often called the mind-body problem. The mind-body problem concerns the relationship between the mind and the body: is the mind nothing more than the physical body with a certain arrangement of parts? Is the mind a collection of causal functions performed by the body? How does consciousness arise from the body? The problem of mental causation is intricately connected in that many answers to the question of how the mind and body are related will have starkly differing consequences for whether or not the mind has any genuine causal efficacy on the body, or on the world via the body. They are, however, different issues. The question of phenomenal consciousness and its relationship to various neurophysiological processes may be answered, for instance, without thereby yielding a firm answer to the question of whether phenomenal consciousness has genuine causal efficacy, on what it can act, etc.

Historical background

The original version of the problem of mental causation as it figures in contemporary debates is often taken to begin with Descartes' Meditations (1641/1996), although it can be dated back as far as Plato's *Phaedo*. Descartes presented a dualist picture of the world as comprised of two distinct substances, physical and mental. Physical substances, or objects made of matter, are spatially extended and located, and are not capable of thought. Anything composed solely of matter moves via mechanical forces only; Descartes thought that animals, for instance, were simply very complicated automata. Mental substances, or minds or souls, are that which thinks; they are not spatially extended or located. Properties had by physical substances are, on Descartes' view, fundamentally incompatible with mental substances. Minds cannot have shapes or motion, and material objects cannot have thoughts or sensations.

Humans are, according to Descartes, an intimate union of a material body that moves like a machine with a mental substance that thinks, that senses via the material body, and that can influence the motion of that material body. Descartes was careful to avoid a view in which the mind and body were overly separate. The mind/body union is not like the ship in which a sailor directs the motion, he says. A sailor can only know that there is damage to the ship's hull by going and inspecting it. But we can know things about our body in a direct way, by sensing, in a way that we cannot with any other material body to which our thinking minds are not intimately connected. We don't control our body in the distant way in which we control a puppet; there is an immediacy to how we move our limbs that is the consequence of our minds being connected in this special way to our body and not to other pieces of matter.

Descartes thus denies physicalism about the mind-body relationship: the mind could not be made out of particular bits of matter; no amount or organization of matter could ever comprise a mind, because it would always be of the wrong kind of substance to do so. But the mind is intimately connected to some particular chunk of matter, to its body. It receives sensations from that body and, most relevant for our purposes, it also moves that body directly, but no other bodies directly. There are then two directions of influence that pose a problem: how does the physical world have a causal effect on the mind via perception? And, how does the mind influences the body?

Descartes' account is illuminating of the trajectory of discourse on mental causation not only because he runs into a certain kind of problem, but also because he then proposes a certain kind of solution that runs into another very characteristic problem. His dualist account generates a clear case of a problem for mental causation by conceiving of the mind as an entirely different kind of thing than the physical body on which it acts. If physical motion is entirely mechanical, and the mind is entirely unphysical, how does the mind communicate motion to the body? His solution is to propose a way-station where physical bumpings and pullings are translated into signals that influence the mind. Sensation involves physical causation that is transmitted via a complicated series of physical machinery to a special place in the brain – famously, he proposed the pineal gland as a possible site for this – that translates the physical signal into a mental one. Once the mind resolves to do something, such as lift an arm, that mental signal is then translated back via the same site to a physical signal that would pull and push on the right parts of the body to make the arm go up.

This solution to how the mind and body causally influence one another doesn't really solve the problem, though. Instead, it relocates it. Proposing a way-station that translates mental and physical influence back and forth between the two kinds of substance physically isolates the location whereby the mind controls the body, but still does not answer the problem of *how*, exactly, the pineal gland translates between two fundamentally different substances. Localizing the physical space in which such a translation happens is arguably an

improvement over allowing it to happen all over the body, but still it does not budge the problem of mental causation: how does that mental influence get to the body, and use that body to influence the world? This has been called by Robb and Heil (2013) a problem concerning the causal nexus by which mind and body are connected: "Any causal relation requires a *nexus*, some interface by means of which cause and effect are connected."

The way in which this problem arises in Descartes' account is particular to his dualist views, and the problem of how the mind could exercise a causal influence in the physical world looks somewhat different when one rejects substance dualism. But it is a problem for causation that will arise in a different form anytime the mental and physical are treated as of different sorts: *how*, not merely *where*, does one causally influence the other?

This problem of how the mental could influence the physical simply does not arise in accounts of the mind that treat it as identical with, or at least of the same kind of metaphysical substance as, the brain – in other words, in any physicalist account of mind. In the 20th century, for instance, the identity theory held that the mind just is the brain – processes in the mind are not merely correlated with, but simply are identical to, processes in the brain (for instance, Smart 1959). In this account, there is no causal influence of the mental beyond that of the causal influence exerted by the brain. Type identity theory is the views that types of mental processes, like pain, simply are types of physical processes, like C-fibers firing. The view that types of mental processes or events are nothing other than types of physical events or processes is also called reductive physicalism: physicalism, because it holds that the mind is physical in character, and reductive, because it holds that the mental reduces to the physical.

While there are different versions of it, reductive physicalism, in any form, solves the tension between a physicalist view of mind and the apparent causal efficacy of the mental by denying that mental causal efficacy is anything other than the causal efficacy of neurophysiological processes. The reductive element of the view means that mental causation is a placeholder for the real causal story which inevitably involves brain processes, microphysical particle states, or some other straightforwardly physical cause. Reductive physicalism about the mind follows a trend in the 20th century that characterizes many scientifically-informed philosophical accounts. It organizes the relationship between types like the mental and physical in terms of levels, such that the causal efficacy (and, indeed, other features like metaphysical fundamentality) of higher levels depend on and reduce to that of lower levels. The mental is causally efficacious only insofar as it reduces to the physical, and the physical is causally efficacious. Reductive physicalists allow that mental terms are a pragmatic convenience and don't need to be eliminated, so long as they are understood as a kind of short-hand for the 'real' causal story. This means that reductive physicalism is less strong than eliminative physicalism, which holds that mental terms will, or at least should, be replaced entirely (P.S. Churchland 1986; P.M. Churchland 1981).

Anomalous monism and causal exclusion

The tension between physicalism and the causal efficacy of the mental was given an influential analysis by Donald Davidson (1980/2001; also, 1995). He aimed to both preserve physicalism as a view about the nature of the mind as part of the physical world studied by science, while also arguing for autonomy from physical causation for causes and causal relations that involve mental terms. The mental was part of the physical world, but could not be reduced to the laws of physics. His view, anomalous monism, was pitched at least partially as an alternative to reductive physicalism. His account offers a nonreductive physicalist way to accommodate both of the intuitions that there is nothing non-material about mental events and that mental events do have genuine causal efficacy.

This section explores Davidson's account and Kim's challenge to it, and how this sets the framework for much of the contemporary debate about mental causation, as well as connecting that debate to the broader issue of higher versus lower level causation generally.

Davidson subscribes to what he calls the Cause Law thesis (1995), which is one of the key premises in his account and closely related to the deductive-nomological model of explanation in science. According to the Cause Law thesis, all true causal statements are such that some general causal law connects the cause and the effect under some description. Singular causal claims are made true because there is an underlying general law of the form, All Xs cause Ys, such that the cause and the effect are instances of Xs and Ys. The singular claim, "that rock just broke this window," might be true because "All rocks, thrown sufficiently hard, break windows," is true. This is where the nomological force of causal relationships comes from – the cause necessitates the effect because of the law connecting them, and without such a law, there would be no connection between cause and effect.

According to Davidson, causes and effects are those things out there in the world, to which we point with our descriptions. As such, Davidson holds that the same causes and effects can be picked out using different descriptions. When we make a true causal claim that x caused y, it is true because of that covering law, regardless of whether the covering law involves x and y so described or if it involves x and y under an entirely different description. Thus, the Cause Law thesis is an existence claim about there being a law under *some* description of the cause and effect, not that there is a law for every description, nor that there is a law that we know, or even that we know the description under which the law holds. It is merely the claim that, whether or not we ever know it, there is such a description and such a law, and it is this which makes true any causal claim, including but not limited to those involving mental causes.

Putting these pieces together, Davidson holds that mental causes (and effects, although mental causation is primarily taken to be problematic for mental causes of physical events) are, extensionally, part of the same physical causal nexus that is studied by the sciences, and

are thus causal because they figure in true general laws. However, there are no laws containing mental terms – no laws in which both X and Y are mental, nor laws in which either X or Y are mental. Mental causes can be genuinely causal, but mental causes and effects are anomalous because it is never as mental that they are covered by a law. Mental descriptions are one way of describing the causes and effects in questions, but not the only way. If one were to redescribe those same causes and effects in other terms, namely, in the right physical terms, then a general law of the form All Xs are Ys would cover every instance of true mental causation. But it would never cover it under the mental description.

In this way, Davidson hopes to salvage features of the mental, like rational constraints on action, while still situating mental causation within the same causal nexus as the causes and effects studies in the sciences. The kinds of causal relationships we find in physics cannot account for or accommodate what he called "the uneliminably normative or rational aspect of intentional idioms, and the consequent irreducibility of mental concepts to concepts amenable to inclusion in a closed system of laws" (Davidson, 1995). Because the mental is anomalous, with no genuine laws involving mental terms, it cannot be reduced to the merely physical. Mental descriptions continue to be useful as a separate domain of discourse, even while acknowledging that the mental cause and effects in question are not anything above and beyond the physical causal nexus with which we are scientifically comfortable. The mental is simply another way, a nonreducible way, of describing certain parts of that nexus.

There are many issues that could be raised (and have been, by a wide variety of authors) with respect to Davidson's account. For instance, it relies on the rather antiquated deductive-nomological account of explanation, where causal explanations must all be made true by universal laws. Given other accounts of explanation and/or causation, it is not clear that anomalous monism could even be formulated, or that mental causation actually poses any kind of particular problem – we'll explore this line of thought more in a subsequent section. Furthermore, it is something of a matter of faith that there really are multiple legitimate descriptions of the *same* mental causal relata. As we'll also see in a subsequent section, there are many cases where redescribing causal relata actually changes the subject, by changing the causal relationships into which it enters. But for now, it suffices to lay out Davidson's view of anomalous monism as a key position that aimed to accommodate mental causation within the commitments of physicalism, against which the main challenge against genuine mental causation, that of causal exclusion, is targeted.

Davidson's account is the primary target for Jaegwon Kim's (1989, 2000) causal exclusion argument, which is arguably the core of the contemporary problem of mental causation. Kim challenges the idea that Davidson has salvaged genuine causal efficacy for the mental, and argues that it is instead a different form of reductive physicalism. Kim's challenge to anomalous monism can be generalized to pose an issue not just for mental causes, but for any causes that are in some way higher-level with respect to another set of potential causal relata.

Kim's influential work involves both a specific criticism of Davidson's account as well as a broader challenge to any nonreductive account of the relationship between the mental and the physical. One of his main criticisms of Davidson's account of anomalous monism is that it does not actually establish genuine causal efficacy for the mental, and, thus, does not block the reduction of the mental to the physical. That some event has a mental description is causally irrelevant, claims Kim (1989). If it is only under a physical description that an event can be covered by a law, and it is only because of a covering law that an event is causally efficacious, then no mental event is ever causally efficacious because it is mental, but only because it is physical.

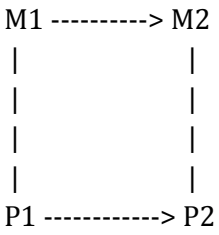
Horgan referred to this as the problem of quausation (1989): mental events enter into causal relationships, but only as physical events, never by dint of their being mental events. The mental is an epiphenomenal ride-along of the physical, meaning that mental causation is merely apparent and nothing other than microphysical causation under a different name. Many authors find this an acceptable conclusion. But, in order to retain physicalism *and* genuine causal efficacy of the mental, something more is needed than simply having any kind of causal efficacy. It needs to be the right sort of causal efficacy, the sort attributable to mentality, not merely physicality.

The broader challenge that Kim has issued is often referred to as the causal exclusion problem. It applies to any nonreductive yet physicalist account of the mental, and sets out what must be overcome in order for mental causation to be genuinely mental, rather than an epiphenomenal redescription of genuine microphysical causation. Kim highlights the tension between physicalism and genuine mental causal efficacy by pointing out that the class of physical causes is supposed to be complete: for any physical effect, there is a sufficient cause of it that is physical, also. This is the rejection of dualism, of 'spooky' nonphysical causation somehow bumping parts of the physical world around. According to anomalous monism, some physical causes are also mental causes. However, the effects of those causes are already 'caused', as it were: for any physical effect, there is already a complete physical causal story.

One is then confronted with the following dilemma: either such effects, with both physical and mental causes, are overdetermined or they are not overdetermined. If they are not overdetermined, then there is something that the mental contributes causally to the effect. But this also violates the assumption of physicalism. On the other horn of the dilemma, however, all effects with mental causes are overdetermined, which means that they have multiple causes each of which is sufficient to bring about the effect. The physical causes alone, with no additional mental cause, would have been sufficient to bring about the effect exactly as it occurred. This horn has two unfortunate consequences. It renders mental causes systematically superfluous, while also committing to a metaphysically suspect view about the incredibly rampant presence of overdetermination in the world for just a particular set of causes, namely, all and only the mental ones. Neither horn of this dilemma

accomplishes the goal of preserving both physicalism and genuine causal efficacy of the mental.

The argument for causal exclusion is often accompanied by the following sort of diagram. M1 and M2 are singular tokens of particular mental events, and P1 and P2 are the corresponding physical event tokens on which the mental events depend.



The vertical lines are of some kind of supervenience or ontological grounding relation between tokens of mental events and tokens of physical events. According to physicalism, all tokens of the mental must also be physical tokens. The horizontal line connecting P1 and P2 is causal: this represents the complete physical cause of P2. The question of mental causation is then the question of the causal efficacy of M1. There is an apparent causal relationship between M1 and M2. But, M1 supervenes on P1, and M2 supervenes on P2. P1 is a cause of P2, which means that there is already a complete cause for M2 in P1. Kim's claim is that M1 reduces to P1 as a cause: there is nothing about M2 that is leftover or uncaused, such that M1 can add by causally contributing to it. M1 does not cause P2, since the completeness of the physical means that P2 already has a sufficient physical cause, P1. What is left for M1 to do, asks Kim? M1 simply reduces to P1 in this picture, which means that the project of salvaging both physicalism and autonomous mental efficacy has failed.

The causal exclusion problem can be generalized to any set of token relata such that one kind of relata is identical with, supervenes on, is instantiated by, or is otherwise dependent on another kind of relata (see, for instance, Shapiro and Sober 2007). In other words, any so-called higher-level phenomena will be subject to similar concerns about their genuine causal efficacy vis-à-vis their lower-level counterparts. Evolutionary fitness, for instance, would, according to the generalized exclusion problem, have no causal efficacy of its own, being merely a stand-in for the causal efficacy of its lower-level instantiations in individual instances of reproductive success. In the case of evolution, this may seem like the appropriate stance to take about the potential causal efficacy of higher-level causes: we should refrain from reifying them all as having some additional mysterious kind of causal efficacy all their own. But in other cases, such as causal efficacy attributed to organisms as reproductive units, rather than, for instances, their genes, this is not as clearly the right result.

The generalized causal exclusion problem is the subject of criticism for the way in which it flattens causation, with the possibility of eliminating it altogether (see Block 2003). If the

causal efficacy of higher level causes depends on or reduces to that of the lower level causes, then we can proceed to ask about the status of those lower level causes. Unless they are metaphysically bedrock, then they are in turn higher level than some other, yet-lower level causes. In this fashion, we proceed down to the microphysical, where we confront a new kind of dilemma. On one hand, we could commit to there being some lowest level such that only events at this level are genuinely causal. On the other hand, if there is no such lowest level, then causation "drains away" (Block 2003) into the bottomless abyss of the microphysical, and there is no genuine causal efficacy to be found anywhere.

The debate about causal drainage is somewhat off-topic for the issue of mental causation, but it helps keep a perspective on the urge to reduce higher level to lower level causes. While this may seem like a clearly justified maneuver in the context of a single pair of causal types (such as mental to physical, or psychological to neurophysiological), it requires making an assumption about what kinds of causes lack genuine efficacy, namely, any that are higher level with respect to other causes. This assumption may undercut the validity of the reduction by rendering those lower level causes just as inefficacious as the higher level ones. If neither the higher nor the lower have real causal efficacy, the impetus to reduce one to the other is greatly diminished.

There is some traction to be gained on the issue of causal exclusion and mental causation by considering the same structural problem for causal efficacy with different types of causes. Philosophers and scientists tend to have deeply entrenched views about what the 'right' answer should be about mental causation, and this makes it easy to construct ad hoc 'solutions' that have little merit other than yielding the correct conclusion, or to reject potentially viable accounts because they get the wrong conclusion. By considering the same question applied to different relata, such as evolutionary fitness or thermodynamic temperature, we can assess different proposals reconciling or rejecting higher-level causal efficacy on more neutral territory, and only then apply the solution to mental causation in particular.

The causal exclusion problem, applied to specifically mental causal relata, is the primary and dominant contemporary problem for mental causation. What is causing what, exactly, when we identify apparently mental causes, or offer causal explanations that rely on causal relationships involving mental relata?

Salvaging mental causation by solving or dissolving the exclusion problem

More solutions for solving the causal exclusion problem have been proposed in the last two decades than are possible to canvas in one place. The responses to this contemporary version of the problem of mental causation can be helpfully categorized in terms of the strategies they employ, however. Considering responses in terms of these strategies is a

useful way to map out the territory of ideas involved in the tangle of mental causation, mind-body relationship, reduction, explanation, and more.

Two popular general strategies are that of (i) changing the metaphysical type of the physical and mental tokens that are the relata in the above diagram, or (ii) changing the relationship posited between those tokens, such that the causal exclusion problem can no longer be formulated. Additional strategies involve various ways to bite the bullet, accepting that there is limited or no causal efficacy to the mental, and either (iii) rejecting the intuition that the mental should be treated as having causal efficacy, or (iv) offering a watered-down form of relevance for the mental that falls short of genuine causal efficacy but offers something more than pure epiphenomenalism. Two additional strategies dissolve the problem, rather than solving it directly: (v) denying the identity of mental and physical tokens, which means denying the relationship represented vertically in the above diagram; and (vi) challenging the implicit assumptions about causation on which the causal exclusion problem relies, with recourse to specific theories of causation. I'll briefly discuss each of these strategies in order.

Strategy (i) starts by noting that the relata of mental and physical tokens used in formulating the causal exclusion problem could be of numerous different metaphysical types. A common approach is to follow Davidson and treat them as singular events, which is also convenient since events are a very common relata type in theories of causation. Token mental events just are specific, very complicated, physical tokens, and the mental event tokens stand in all and only the causal relations in which the physical event tokens stand. Token-identity of events, however, leads straight to the causal exclusion problem whereby mental events do not have genuine causal efficacy. There are multiple potential candidates for the metaphysics of these tokens, but what unites these different accounts is that they each reconfigure the relata in the diagram above, as a way of attempting to block the reduction of the mental to the physical.

Instead of treating the mental and physical relata as events, one could construe them in terms of properties. Perhaps the mental is a second order property, i.e., a property had by some other, first order, property. If the mental is a second order property of first order physical properties, then it does not reduce away – it could be causally relevant or efficacious as a property of properties. There is thus only one token, having both physical, lower level, properties and mental, higher order, properties. Each instance of a mental property and a physical property are tokened by the same object. The mental property is multiply realizable, capable of being instantiated by a wide range of quite distinct physical tokens. As such, the mental properties do not reduce to the physical ones (see, on this, Bennett 2003, Levin 2009).

Strategy (ii) is similar to (i) in attempting to block causal exclusion by changing the characterization of the diagram above, but instead of focusing on the relata, the focus is on the relationships between them. The horizontal lines are causal; but what precisely is

represented by the vertical lines? How one cashes out the relationship between the mental and the physical will affect the reducibility of mental causation to physical causation. A common approach is that of some kind of supervenience of the mental on the physical (see especially Wilson 2005).

Supervenience can be a very broad asymmetric relationship, committed to the claim that there can be no change in the mental without there also being a change in the physical, while there could potentially be a change in the physical without a change in the mental. For global supervenience, the mental parts of the world (be they events, properties, tropes, etc. – see strategy (i)) supervene on the entire physical world, such that any change, no matter how insignificant, might be sufficient for an entirely different set of mental relata to supervene. If we are considering the possible causal efficacy of Alice's intention to surprise Bob on the physical outcome of Bob's startled jump, then it is extremely counterintuitive to commit to the claim that the mental cause has changed because of a slight rearrangement of particles in a distant galaxy. Surely the mental relata supervene on a somewhat smaller batch of the physical world. Picking out what, exactly, goes into the patch of the physical world on which a particular candidate mental cause supervenes has proven to be a contentious issue.

Strategy (iii) involves denying the intuition that there is autonomous mental causation. Reductive physicalism of any stripe is an example of this, of which Kim himself is a well-known proponent. There are many versions of reductive physicalism, where the basic premise is that whatever the mental is, it reduces to or is nothing over and above the physical. Even stronger views, like eliminative physicalism (Churchland 1981), advocate the eventual eschewal of all mental terminology.

Strategy (iv) takes a somewhat different tack. Instead of simply denying the legitimacy of the intuition, some philosophers offer a replacement for genuine causal efficacy, something that explains why it seemed as if the mental were causally efficacious without having to grant full-blown mental causation. Jackson and Pettit (1990), for instance, distinguish between causal efficacy and causal relevance. Causal efficacy is what actually causes things to happen; causal relevance is what properties have such that they are cited in good causal explanations. But, while causally relevant properties may be explanatory, they lack causal efficacy. For example, according to this view, being fragile is causally relevant to the vase breaking, even though it was not the fragility that was causally efficacious in the actual breaking. This applies to the debate regarding mental causation by characterizing mental properties as causally relevant, even while it is always some other, quite specific, set of physical properties that are causally efficacious.

Strategy (v) is related to (i and ii) in that it focuses on the character of the relata and relationships in the diagram above, but instead of changing the kind of relationship between the mental and the physical, it argues that the vertical lines in the diagram don't exist. In other words, it denies the identity of token mental and physical relata, and by doing so,

prevents the causal exclusion problem from being formulated. If the token of a mental event is not identical to or does not supervene solely on the token of a physical event, then the very way in which the problem is framed is specious (Andersen 2009).

There is a delicate balance to be struck in such strategies for responding to the causal exclusion problem for mental causation. The relationship between the mental and physical is itself as much a live issue for discussion, and as unresolved, as the existence or nature of mental causation itself. On the one hand, if we are willing to assume that the mental does have genuine causal efficacy, then we can use the causal exclusion problem as a rough guide towards how best to represent the relationship between the mental and physical. It would serve as a constraint on our representations that they yield the correct answer with respect to mental causation, and this would rule out some construals, such as token-identity of mental and physical events. It would not constrain enough to yield a single unequivocal solution, but it would be a substantive guide. On the other hand, though, we should be legitimately concerned about making this assumption about genuine mental causal efficacy, since we might take the causal exclusion problem to show exactly that this assumption is unwarranted. In that case, it is ad hoc to gerrymander our characterizations of the mental and the physical in order to reach the conclusion that the mental is genuinely efficacious. Rather than guiding us towards solving other problems like that of the mind-body relationship, it begs precisely the question at issue.

This leads to strategy (vi): focus on causation as a way to situate mental causation in a broader perspective with better evidential footing. This tactic allows us to pose the question of whether or not, and in what way, the mental has genuine causal efficacy without having to make assumptions about how to best represent the relationship between the mental and physical relata in question. Accounts of causation provide independent evidential criteria for what it takes to count as a cause of something else; we can use these to investigate whether any mental 'causes' actually meet the criteria for causation.

One way in which this has been done is to challenge the treatment of the mental and physical relata as sufficiently distinct that we can even sensibly ask the question about which one does the causing (Dardis 1993; Campbell 2010; Andersen 2009). Kim's original question, of whether it is M1 or P1 that causes M2, is misleading in that it asks us to implicitly treat these as distinct causal relata. They simply aren't in competition, such that one – the physical – 'wins out' over the other. Arguing from causal exclusion to the inefficacy of the mental is, on this strategy, a kind of category mistake.

Another way that strategy (vi) can be implemented to undermine the causal exclusion problem is to turn to contemporary accounts of causal explanation and see how mental causal relata fare on such accounts. Davidson's assumption about causation requiring general laws is widely rejected, and the causal exclusion problem cannot be formulated within many contemporary accounts of causation. In the last several decades, remarkable progress has been made in developing sophisticated philosophical and mathematical

techniques for analyzing causal structure. We can translate the question of mental causation into specific accounts of causation to see if there is anything genuinely causal about the mental in each account. Once we do this, it turns out that on almost every contemporary account of causation, the mental has as much causal efficacy as any other cause outside of fundamental physics.

Consider an example of this. On the influential interventionist account of causation (Woodward 2003), the question of whether or not the mental has causal efficacy is made more specific by considering a variety of causal variables that could represent different aspects of mental causation. One could ask about the role of conscious visual awareness in actions like reaching for and grasping objects. Each of these would be treated as a variable that takes on different values: perhaps Conscious visual experience (yes, no) and Reaching (hand pre-formed, hand not pre-formed). One would then consider the empirical research on this in order to determine if these two variables meet the criteria for having a causal relationship between them. Put very simply, this is a way of asking whether or not intervening on conscious visual awareness of an object changes the way in which we reach for those objects. Once we establish the answer (which, in this case, is yes – see Andersen 2009, chapter 3), we've shown that there is at least one case of mental causation, namely that which is represented by the variables in question.

Against this kind of mathematically sophisticated and scientifically grounded account, Kim's causal exclusion argument looks rather simplistic and naïve. One might want to protest that the variables are just a stand-in for the real causal story, which surely involves just the neurophysiological processes initiated by light impinging on the retina, and so forth. To defend this response in the face of a well-validated account of causation that has independent justification for its requirements on what counts as a cause, one should have more than just the intuition that the 'real' causal story is elsewhere. This is a key advantage to treating the problem of mental causation as one of causation in general, rather than one of mental causes as sui generis: given such well-established and explicitly justified methods for finding causal structure, it takes a great deal to show that the answer yielded by accounts of causation such as interventionism are incorrect. This puts the onus on reductive physicalists to show what is wrong in such analyses and to offer a more substantive defense of what counts as a 'real' causal story, a very challenging task.

These six different argumentative strategies cover a vast number of different accounts, organized in terms of commonalities they share with regard to the aspect of the causal exclusion problem that they address. While there are multiple aspects to the issue of mental causation, there is no doubt that this problem, reconciling genuine mental efficacy with a physicalist view of the world and an ambiguous relationship between the mental and physical, is one of the main debates in contemporary philosophy of mind.

Directions for future research on mental causation

The causal exclusion problem is interesting for the way in which it captures so clearly the dilemma for genuine mental causation. The generalized exclusion problem raises issues for many different kinds of higher level causes, including the mental as well as a huge host of other potential causal relata from various sciences. However, the version of causal exclusion that applies to the issue of mental causation has, as we've seen, a unique twist that renders it particularly difficult to deal with. There is no clear answer to the question of how the levels in question are related to one another.

It is often tempting to treat mental causes as simply stand-ins for the 'real' causal story, which must be purely physical and involve neurophysiological processes. After the last section, we should be cautious about this practice. Once we meet the evidential requirements for independently-justified accounts of causation for the mental to have genuine causal efficacy, it is unnecessary to reject mental causation by claiming that 'really' what we've just shown is somehow false or merely apparent. Furthermore, it is a substantive task to translate one potential relatum into another. Moving too quickly from the mental to the neurophysiological risks changing the subject, where the new physical process or event may be causally efficacious, but not efficacious *of the same effect* as was the original mental relatum.

This is where much of the future work regarding mental causation may be directed: towards a careful and detailed elaboration of the variety of ways in which mental causal relata can be translated into a format such that accounts of causation, coupled with results from cognitive neuroscience and psychology, can yield specific answers. The causal exclusion problem pitches mental causation as an all or nothing problem: either it's causal or it's not. Moving away from this approach means moving towards a scientifically enriched process using well-confirmed tools from studies of causation. The goal then becomes to suss out the structure of mental causation and how it is situated in larger structures of the body and the environment.

Davidson was onto something when he said that mental causation, whatever it was, needed to be the same ordinary sort of causation that is studied in the sciences. It cannot be something different or new without thereby bringing in the sorts of mysterious powers that anyone with a commitment to physicalism of any stripe should want to avoid. He was wrong about what that kind of causation is – there is no need to assume unknowable universal laws governing physical redescriptions of mental causal relata. The way forward in this debate will be through foregrounding the issue of causation, making the question of mental causation more explicitly one of mental *causation*.

References

Andersen, H. (2009). *The Causal Structure of Conscious Agency*. University of Pittsburgh: http://d-scholarship.pitt.edu/9254/

Beckermann, Ansgar (1992). Reductive and nonreductive physicalism. In Beckermann, Flohr & Kim (eds.), *Emergence or Reduction?: Prospects for Nonreductive Physicalism*. De Gruyter; 1-21.

Bennett, K. (2003). Why the Exclusion Problem Seems Intractable and How, Just Maybe, to Tract It. *Noûs* 37(3), 471-97.

Block, N. (2003). Do Causal Powers Drain Away?. *Philosophy and Phenomenological Research,* 67(1), 133–150

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues* 20(1), 64-79

Churchland, P. M., (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78, 67–90.

Churchland, P.S., (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.

Dardis, A. (1993). Sunburn: Independence Conditions on Causal Relevance. *Philosophy and Phenomenological Research,* 53 (3), 577-598.

Davidson, D. (1980/2001). *Essays on Actions and Events* Oxford: Oxford University Press.

-- (1995). Laws and Cause. *Dialectica* 49(2-4), 263-280.

Descartes, R. (1641 / 1996) *Meditations on the First Philosophy*, J. Cottingham (trans.), Cambridge: Cambridge University Press.

Horgan, T. (1989). Mental Quausation. *Philosophical Perspectives* , 3, Philosophy of Mind and Action Theory, 47-76

Jackson, F., and Pettit, P. (1990). Program Explanation: A General Perspective. *Analysis*, 50(2), 107-117.

Kim, J. (1989). The Myth of Nonreductive Materialism. *Proceedings and Addresses of the American Philosophical Association*, 63(3), 31-47.

-- (2000). *Mind in a Physical World* Cambridge, MA: MIT Press.

Levin, J. (2010). Functionalism. *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/functionalism/>.

Robb, D. and Heil, J. (2013). Mental Causation. *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/mental-causation/>.

Shapiro, L., and Sober, E. (2007). Epiphenomenalism: The dos and don'ts. In *Thinking about causes: From Greek philosophy to modern physics*, Edited by: Wolters and Machamer, 235–264. University of Pittsburgh Press.

Smart, J.J.C (1959). Sensations and Brain Processes. *Philosophical Review,* 141-156.

Wilson, J. (2005). Supervenience-based formulations of physicalism. *Nous,* 39(3), 426-459.

Cross-references

Consciousness and agency
Determinism and the Brain
Explanation and Levels
Freewill, Agency, and the Brain
Realization, Reduction, and Emergence