Erkenntnis (2022) 87:53–73
https://doi.org/10.1007/s10670-019-00184-5

ORIGINAL RESEARCH

Mental Causation, Autonomy and Action Theory

Dwayne Moore[1]

Received: 6 March 2019 / Accepted: 3 October 2019 / Published online: 28 October 2019

Abstract
Nonreductive physicalism states that actions have sufficient physical causes and distinct mental causes. Nonreductive physicalism has recently faced the exclusion problem, according to which the single sufficient physical cause excludes the mental causes from causal efficacy. Autonomists respond by stating that while mental-to-physical causation fails, mental-to-mental causation persists. Several recent philosophers establish this autonomy result via similar models of causation (Pernu, Erkenntnis 81(5):1031–1049, 2016; Zhong, J Philos 111(7):341–360, 2014). In this paper I argue that both of these autonomist models fail on account of the problem of Edwards's Dictum. However, I appeal to foundational principles of action theory to resuscitate mental-to-mental causation in a manner that is consistent with the models of causation endorsed by these autonomists.

Mental causation seems true: Marvin is crying because he is sad that his gecko died, Jenny eats a peach because she wants one, *etcetera*. Physical causal completeness seems true as well: Marvin is crying because a complex physical process caused his lacrimal sac to secrete tears onto his eye surface. Hence nonreductive physicalism: Marvin is crying because he is sad and because of the complex physical process. But nonreductive physicalism faces the exclusion problem: events can have no more than the single sufficient physical cause, so mental causes are excluded from efficacy. Autonomists respond by agreeing that mental-to-physical causation fails, but mental-to-mental causation persists: Marvin is crying because of a complex physical process, but Marvin is sad because he believes his gecko died. Several recent philosophers establish this autonomy result via similar models of causation (Pernu 2016, Zhong 2014). In this paper I argue that both of these autonomist models fail on account of the problem of Edwards's Dictum. However, I appeal to foundational principles of action theory to resuscitate mental-to-mental causation in a manner that is consistent with the models of causation endorsed by these autonomists.

✉ Dwayne Moore
dwayne.moore@usask.ca

[1] Philosophy Department, University of Saskatchewan, 9 Campus Drive, Saskatoon, SK S7N 5A5, Canada

The paper is divided into five sections. First, I outline the exclusion problem (Sect. 1), followed by the autonomy solution as recently propounded by Zhong (2014) and Pernu (2016) (Sect. 2). I then outline how the problem of Edwards's Dictum undermines their autonomist solutions (Sect. 3). I then argue that foundational principles of action theory conflict with Edwards's Dictum (Sect. 4), before re-establishing mental-to-mental causation and considering objections (Sect. 5).

1 The Causal Exclusion Problem

The causal exclusion problem is the conjunction of several individually plausible, but (seemingly) jointly inconsistent principles. Here are the first two principles:

> *Physical Causal Completeness*: "If a physical event has a cause at *t*, it has a sufficient physical cause at *t*" (Kim 2009, 38).
> *Supervenience*: "Whenever something has a mental property, *M*, at *t*, it does so in virtue of the fact that it has, at *t*, a physical base property, *P*, where *P* necessitates *M*" (Kim 2009, 40).

Physical causal completeness, as supported by various conservation laws (Montero 2003; Papineau 2001 13ff) and various successes in the neurosciences (Papineau 2001, 31; Melnyk 2003, 238ff), is often taken to be "fully established" (Papineau 2001, 33). Supervenience is intuitive in numerous instances: the object determines the shadow, the apple's redness determines that the apple is coloured, the horse-wise arrangement of parts determines that there is a horse. Combining supervenience with physical causal completeness yields the result that all events and properties that occur at *t* have a sufficient physical cause and/or are determined by physical properties of events. For the purposes of this paper, call the conjunction of these two principles the doctrine of physicalism.

Physicalism comes in several varieties, where reductive physicalism and nonreductive physicalism are chief among the options. Nonreductive physicalists typically supplement physicalism with the following two principles:

> *Mental Causation*: mental properties of events sometimes cause mental effects and physical effects.[1]
> *Irreducibility*: mental properties of events are non-identical with physical properties of events.

Mental causation is supported by appeals to common sense intuitions, as well as argumentation suggesting that mental causation is needed as a foundation for the moral responsibility of agents. For these reasons, most agree with Jerry Fodor that

[1] Some frame the principles of mental causation and irreducibility in terms of mental events and physical events (Kim 2005, 42), while others frame these principles in terms of mental properties of events and physical properties of events (Fodor 1974, 100). I will frame these principles in terms of mental properties of events and physical properties of events, but nothing of substance rides on this distinction.

abandoning mental causation would amount to "the end of the world" (Fodor 1989, 77). Irreducibility is supported by the multiple realizability of mental properties (Putnam 1967): Jane's hunger can be realized by some disjunction of physical properties [$p \vee p_2 \vee p_3 \vee \ldots p_n$]. Multiple realizability, combined with intuitive distinctions between mental and physical properties (Chalmers 1996; Lowe 2006, 15–16), lead many to find the principle of irreducibility "obviously true" (Bogardus 2013, 446). Combining these two principles yields the result that mental properties of events, as non-identical with physical properties of events, sometimes cause mental effects and physical effects. Call this irreducible mental causation.[2]

Physicalism, yoked with irreducible mental causation, yields the following result. From physical causal completeness, the physical properties p of brain events are sufficient causes of the physical properties p^* of bodily movements. And, from the combination of mental causation and irreducibility, the physical properties p^* of bodily movements are also caused by mental properties m of brain events. Thus, some physical properties p^* of bodily movements have both a sufficient physical cause p and a mental cause m. Call this Physical Effect Overdetermination:

> *Physical Effect Overdetermination*: some physical effects p^* have both a sufficient physical cause p and a mental cause m.

The leading objection to this nonreductive physicalist view is the causal exclusion problem, which suggests that physical effect overdetermination is implausible. The causal exclusion problem, as popularized by Jaegwon Kim, rests on the following principle:

> *Causal Exclusion*: "No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination" (Kim 2005, 42).[3]

Here is how the causal exclusion principle poses problems for the nonreductive physicalist doctrine of physical effect overdetermination. Nonreductive physicalism posits that p^* has a sufficient physical cause p and a mental cause m. The causal exclusion principles states that p^* cannot have both p as a sufficient physical cause and m as a cause as well. The result: either p^* does not have a sufficient physical cause p, which violates the nonreductive physicalists adherence to physicalism, or p^* does not have a mental cause m, which violates the nonreductive physicalists adherence to irreducible mental causation. Either way, nonreductive physicalism fails.

[2] Irreducible mental causation is contrasted with reduced mental causation, where reduced mental causation is the conjunction of mental causation and the rejection of irreducibility: mental causes, as identical with physical causes, sometimes cause mental effects and physical effects. Irreducible mental causation, if possible, is preferable to reduced mental causation (cp. Silberstein 2001, 84; Lowe 1993, 631–632; McGinn 1989, 137). Indeed, even Jaegwon Kim, who ultimately endorses reduced mental causation, worries that reduced mental causation seems like "a form of mental irrealism" (Kim 1998, 199) that "all but banishes the very mentality it was out to save" (Kim 1995, 194).

[3] Genuinely overdetermined events are events caused by two independent causal processes. Sally and Ted both throw rocks, which both smash the window at the same time. This is a case of genuine overdetermination, as these throws are independent of one another, and individually sufficient for the effect. By supervenience, however, m is not independent of p, so the causal exclusion principle does not permit p^* to have both m and p as causes.

2 The Autonomy Solution

While numerous responses to the causal exclusion problem exist, of present interest is that group of philosophers who resolve the causal exclusion problem by appealing to the autonomy solution (Pernu 2016; Zhong 2014; Moore 2013, 836; Nagasawa 2010, 42; Cavedon-Taylor 2008, 106; Gibbons 2006, 82; Crisp and Warfield 2001, 306–307; Marras 1998, 448–449; Tuomela 1998, 22–23; Thomasson 1998, 188; Burge 1993, 116; Jackson 1982, 133). In order to properly define the autonomy solution, the principle of mental causation must be divided into its two conjuncts:

> *Mental-to-Mental Causation*: mental properties of events sometimes cause mental effects.
> *Mental-to-Physical Causation*: mental properties of events sometimes cause physical effects.

Where the principle of mental causation listed above stipulates that mental properties of events sometimes cause mental and physical effects, these two conjuncts divide this principle into two. Mental-to-mental causation occurs when mental properties of events cause mental effects, and mental-to-physical causation occurs when mental properties of events cause physical effects.

The autonomy solution avoids the causal exclusion problem by abandoning mental-to-physical causation but embracing mental-to-mental causation.[4] Mental properties of events do not cause physical effects, so p^* has only a sufficient physical cause p, so causal exclusion is not violated and physical causal completeness is endorsed. But since mental properties of events cause mental effects, irreducibility and a weakened but serviceable version of mental causation is also endorsed—not a bad haul.

While numerous authors propose the autonomy solution, I want to focus on two recent philosophers who deploy similar models of causation to arrive at the autonomy result (Zhong 2014; Pernu 2016). According to them, m causes m^* if both a presence condition and an absence condition on causation are satisfied (Zhong 2014, 345), where these conditions can be discerned via the following counterfactuals (Pernu 2016, 1043):

(1) $m \mathrel{\square\!\!\rightarrow} m^*$

[4] Some autonomists, including Zhong, provide additional nuance here. He agrees that mental-to-mental causation is true, and mental-to-physical causation is false, but he also endorses mental-to-higher-level property causation, which he thinks includes behavioural and social properties (Zhong 2014, 350). This position, however, still faces the same Edwards's Dictum challenge discussed in Sect. 3 that the autonomist endorsing only mental-to-mental causation faces. Namely, for Zhong, higher-level behavioural properties would be completely determined by their lower-level subvenient physical bases, which would exclude mental properties of events from causally interacting with higher-level behavioural properties. For the sake of brevity, I will only discuss the autonomist who endorses only mental-to-mental causation, though the results apply more broadly.

(2) $\sim m \mathrel{\Box\!\!\rightarrow} \sim m^*$[5]

Counterfactual (1) is true, since the nearest worlds where m is present are worlds where some p is present (by supervenience), so some p^* is present (by completeness), which determines that m^* is present (by supervenience).[6] Counterfactual (2) is true, since the nearest worlds where m is absent are worlds where no subvening base p occurs (by supervenience), hence no physical event p^* occurs (by completeness), determining that m^* is absent (by supervenience) (Zhong 2014, 356; Pernu 2016, 1043).

[5] In greater detail, Lei Zhong's autonomy solution appeals to what he calls the "dual condition conception of causation" (Zhong 2014, 342), which is rooted in Chris Woodward's interventionist model of causation and bares affinity with the difference-making model of causation presented by List and Menzies (2009). According to Zhong's view, X is a cause of Y if the following two conditions are true: (A) If an intervention that sets $X=x_p$ were to occur (while all other relevant variables in the causal graph are fixed), then $Y=y_p$; (B) If an intervention that sets $X=x_a$ were to occur (while all other relevant variables in the causal graph are fixed), then $Y=y_a$ (Zhong 2014, 344). The variable x_p in (A) stands for x being present, so (A) roughly states that if an intervention that makes x present were to occur, then y is present. As such, (A) is called the Presence Condition (Zhong 2014, 345). The variable x_a in (B) stands for x being absent, so (B) roughly states that if an intervention that makes x absent were to occur, then y is absent. As such (B) is called the Absence Condition (Zhong 2014, 345). Thus, these counterfactuals adequately approximate Zhong's Absence Condition and Presence Condition, as applied to mental causation (cp. Zhong 2014, 356). It is worth noting that Zhong has recently modified his position. He now rejects the principle of physical causal completeness, thereby allowing him to claim that mental causes bring about physical effects without violating causal exclusion (Zhong 2019). Tuomas Pernu appeals to a modified version of the counterfactual model of causation, according to which X is a cause of Y if the following two counterfactuals are true: (A) Had X occurred, then Y would have occurred ($X \mathrel{\Box\!\!\rightarrow} Y$); ($B$) Had X not occurred, then Y would not have occurred ($\sim X \mathrel{\Box\!\!\rightarrow} \sim Y$). The traditional counterfactual model, as popularized by David Lewis (1973), takes (A) to be automatically true, since the nearest possible world where X is true is the actual world in which X and Y occur, so Pernu's (B) is of central concern. Pernu does not follow this strongly centred view, according to which there is one nearest possible world which is the actual world. Rather, he endorses a weakly centred view, according to which there may be many nearest possible worlds, some of which are not the actual world (Pernu 2016, 1037–1040). This model, which resembles the difference-making model of causation endorsed by Christian List and Peter Menzies (List and Menzies 2009, 483–484), does not result in the automatic truth of (A). Thus, both (A) and (B) must be established as true in order to determine causation. Thus, these are the germane counterfactuals to discern mental causation.

[6] Establishing what the nearest possible worlds are requires a similarity metric for possible worlds. Neither Zhong nor Pernu offer explicit details about their similarity metric. The following details are available, however. Pernu operates within Lewis' counterfactual model of causation, while only diverging with Lewis' strong centring view that the nearest possible world is the actual world. Lewis's similarity metric procedes as follows: the nearest worlds (besides the actual world) are the worlds where a local violation of law altars a particular fact, further worlds are worlds with larger divergences of particular facts across space and time, and the furthest worlds are worlds with widespread violation of law (Lewis 1986, 47). Zhong also rejects Lewis's strong centring view that the nearest world is the actual world, so $m \mathrel{\Box\!\!\rightarrow} p^*$ does not come out as automatically true for Zhong either (Zhong 2014, 345). Beyond this, while Zhong operates within Lewis' counterfactual model in other papers (Zhong 2011), he operates within Woodward's interventionist model in the target paper. Woodward's interventionist model has both similarities and differences with Lewis' similarity metric of possible worlds (Woodward 2003, 133–145). For Woodward, an intervention that breaks a causal chain is similar to a local violation of law that altars a particular fact. So, in some ways Woodward's model is extremely strict in following Lewis's similarity metric. That is, on interventionism, it is local violations of law that occur, even if these local violations of law generate large divergences of particular facts across spacetime, or conjoin to form a widespread violation of law.

With respect to mental-to-physical causation, the presence and absence conditions can be discerned via the following two counterfactuals:

(3) $m \mathrel{\square\!\!\rightarrow} p^*$
(4) $\sim m \mathrel{\square\!\!\rightarrow} \sim p^*$

Counterfactual (4) is true, since the nearest worlds where m is absent are worlds where m^* is absent [by (2)], which are also worlds where no subvening base p^* occurs (by supervenience). But counterfactual (3) is false, since the nearest worlds where m is present are worlds where m^* is present [by (1)]. But, m^* is multiply realizable (by irreducibility), so m^* can occur while p^* is absent but replaced by some slightly different physical realizer p_2^* (Zhong 2014, 355–356; Pernu 2016, 1041).[7] The result: m causes m^* while m does not cause p^*, which is the autonomy view. As Zhong summarizes: "… even if m causes m^*, m may not cause p^* … Equivalently, even if m does not cause p^*, m could still cause m^*" (Zhong 2014, 357).[8] Or, as Pernu concludes: "… neither does the mental causally influence the physical (realisers of m^*) … However, and contrary to initial intuitions, these results do not force one to conclude that the mental is epiphenomenal, for the mental can be shown to have a specific causal role, namely it can be shown to be efficacious in its own realm" (Pernu 2016, 1043).

3 Edwards's Dictum and Supervenience

In this section I introduce the problem of Edwards's Dictum, which is a troubling objection to the autonomist view. To see the problem, recall how the autonomy solution was motivated. Nonreductive physicalism posits irreducible mental causation and physical causal completeness. The result: physical effect overdetermination, as p^* has both a sufficient physical cause p and a mental cause m. The problem: p^*'s sufficient physical cause p excludes the mental cause m. The autonomist solution: p^* has only a sufficient physical cause p, whilst m causes m^*, so m is not excluded from causal efficacy *tout court*.

[7] Pernu and Zhong combine the absence of p^* with the presence of some other realizer p_2^*. This is controversial. First, some operating within a counterfactual model argue that the absence of p^* should not be combined with the presence of some other realizer p_2^*, but should be a clean excision without replacement by p_2^* (Harbecke 2014, 366; Bennett 2003, 482). On this excision model, the absence of p^*, without replacement by another realizer, amounts to the absence of m^* as well, by supervenience. This result would damage Pernu's and Zhong's claim that the presence of m also makes m^* present, thereby calling mental-to-mental causation into doubt. This is a problem within the interventionist model that Zhong deploys as well. Zhong's two conditions on causation both require holding other variables fixed (Woodward 2008, 240), yet his solution sets a third variable p_2^* from absent to present, which is not to hold other variables fixed (cp. McDonnell 2017, 1467). That being said, I grant Pernu and Zhong the view that the presence of m may lead to the presence of m^* being realized by p_2^*.

[8] In this quotation, as well as several others, I modify the names of the variables to retain consistency throughout this paper.

Unfortunately, a similar line of reasoning problematizes the autonomist view that *m* causes *m**. From irreducible mental causation, *m** has a mental cause *m*. From supervenience, *m** is completely determined by subvening physical properties *p**. Thus, some mental properties *m** of bodily movements are both completely determined by subvening physical properties *p** of the bodily movements and are caused by mental properties *m* of brain events. Call this Mental Effect Overdetermination:

> *Mental Effect Overdetermination*: Some mental effects *m** are both completely determined by some physical base *p** and have a mental causes *m*.

It is tempting to deploy the causal exclusion principle against mental effect overdetermination. Here is how that looks: no single effect can have more than a single sufficient cause occurring at a time, so the mental effect *m** cannot both have both a mental cause *m* and be completely determined by *p**. Assuming that supervenience prevails, *m** is completely determined by *p**, so *m* is excluded from causing *m**. However, this is a strained application of the causal exclusion principle. Orthodoxy is that *p**, while determining *m**, does not actually cause *p** (Kim 2005, 36; 1999, 32; Engelhardt 2017). Causes are metaphysically independent and temporally diachronic from their effects, while supervening properties are metaphysically dependent on, and temporally synchronic with, their determining bases (Hume 1888, 78–80; Craver and Bechtel 2007), so supervenient relations are not causal relations. Thus, there is no formal tension between the causal exclusion principle and the view that *m** has a (sufficient) mental cause *m* and is completely determined by *p**. In other words, it is possible to agree that events cannot have more than a single sufficient cause, while also arguing that *m** has *m* as a sufficient cause, while it is also non-causally determined by *p** (Campbell 2015, 58–59; Moore 2013, 835–836; Marras 2007, 309–310; Gibbons 2006, 89; Jacob 2002, 651; Thomasson 1998, 183–184).

While the causal exclusion principle may not pose difficulties for mental effect overdetermination, it is not difficult to reframe the exclusion principle in a way that does pose difficulties:

> *Principle of Determinative/Generative Exclusion*: "If the occurrence of an event *e*, or an instantiation of a property *P*, is determined/generated by an event *c*—causally or otherwise—then *e*'s occurrence is not determined/generated by any event wholly distinct from or independent of *c*—unless this is a genuine case of overdetermination" (Kim 2005, 17).

The determinative exclusion principle is a broader version of the causal exclusion principle that only allows effects to have one complete causal or non-causal determinative source. This principle is consistent with the term 'overdetermination', which also does not simply emphasize over-*causation*, but rather emphasizes the broader category of over-*determination*. According to this principle, *m** cannot be determined by two determinative processes, so either the causal process from *m* to *m**, or the determinative process from *p** to *m**, must be excluded.

The question becomes: who wins the conflict between *m* causing *m** and *p** completely determining *m**? Jaegwon Kim uses Edwards's Dictum to show that that the determinative relation from *p** to *m** excludes the purported causal relation from *m* to *m**:

> *Edwards's Dictum*: There is a tension between 'vertical' determination and 'horizontal' causation. In fact, vertical determination excludes horizontal causation" (Kim 2005, 36)

Edwards's Dictum is named after the American philosopher-theologian Jonathan Edwards, who argues that the universe is re-created every moment by God. This possibility enables us to think of the moments of the universe as disconnected instants, where within each moment, the lower-level events fully determine higher level events, rendering previous moments irrelevant to how events are determined at an instant. For example, the many individually etched lines that exist on the page at one instant fully determine that there is a picture of Bugs Bunny eating a carrot on the page at that instant, no matter what happened in the previous still frame pictures, and no matter who drew the picture some time ago.

While Kim does not endorse Edwards's theology, he argues that the principle of supervenience provides two of the three foundational components of Edwards's Dictum. First, according to Supervenience, *p** is sufficient for *m**. The subvening base *p** at *t* is "fully sufficient" (Kim 2005, 37) for *m** at *t*, or "necessarily sufficient for *m** to be instantiated at *t*" (Kim 2009, 40). Thus, second, consistent with Supervenience, *m* is not necessary as a cause of *m**. Since *p** is fully sufficient for *m** at *t*, *m** will be present at *t* "no matter what happened before *t*" (Kim 2009, 40). Thus, what happened before *t* "seems irrelevant" (Kim 2005, 37), since "as long as *p** is there, *m** will be there *even if the m*-instance's putative cause, the m-instance, had not been there at all*" (Kim 2009, 40; cp. Kim 2005, 37). Since supervenience indicates that *p** is a sufficient determinant of *m**, this also suggests that *m* is not necessary as a cause of *m**—*m** will still be present, even if *m* doesn't occur at all! While Edwards's Dictum takes the third step of suggesting that *m* is thereby excluded from causing *m**, supervenience alone "puts the causal status of the *m*-instance vis-à-vis the *m**-instance in jeopardy" (Kim 2009, 40). For Kim, his supervenience argument against nonreductive physicalism is the fact that supervenience alone indicates that the vertical *p** to *m** determination relation jeopardizes the horizontal *m* to *m** causal relation (Kim 2009, 40). This issue is "at the heart of the worries about mental causation" (Kim 2005, 38), and is the "fundamental idea" (Kim 2005, 36; cp. Kim 1998, 37) of the exclusion problem.[9]

[9] Indeed, not only does Kim use Edwards's Dictum to exclude direct mental-to-mental causation, but he then uses Edwards's Dictum to prove that any instance of mental-to-mental causation must proceed via downward causation. According to Edwards's Dictum, the only way for *m** to occur is for *p** to determine *m**, so the only way *m* can cause *m** is for *m* to cause *p** to determine *m**. At first glance, this seems odd—Kim says mental properties cannot have mental or physical causes (by Edwards's Dictum), but mental properties can cause mental or physical effects (by Downward Causation). The oddity is quickly dissolved, however, as Kim only introduces the possibility of *m* downwardly causing *p** in order to reject this possibility since *p* is the sufficient cause of *p**. As discussed in Sect. 1, it is this downward causation requirement that immediately leads to physical effect overdetermination, so Kim uses Edwards's Dictum to generate multiple problems for mental causation.

Edwards's Dictum foists the following difficulty on Zhong and Pernu's model of mental-to-mental causation. Zhong and Pernu agree that m^* supervenes on p^* at time t. According to Edwards's Dictum and supervenience more broadly, this means that m^* is fully determined by p^*, which renders m unnecessary as a cause of m^*, which jeopardizes the claim that m causes m^*, which undermines the viability of mental-to-mental causation, thereby undermining the autonomy solution.

4 The Causal Theory of Action

The result of Sect. 3 is that mental-to-mental causation is jeopardized by Edwards's Dictum and supervenience more broadly. In this section I show how the causal theory of action strengthens the requirement for mental-to-mental causation.

I want to start some distance off, with an uncontroversial saying: there are varieties of mental properties of brain events, including percepts, beliefs, memories, desires, emotions, moods, actions, *etcetera*. Some of these mental properties seem to be necessarily connected with other mental properties: it seems to be in the nature of 'being a memory' to be necessarily connected to the originating percept, it seems to be in the nature of 'being a conviction' to be necessarily connected to originating beliefs, and, most importantly for present purposes, it seems to be in the nature of 'being an action' to be necessarily connected with reasons for the action. With respect to this last category, it is worth noting that actions themselves come in various sorts. There are internal actions: Jane is trying to focus on the lecture. There are actions without visible bodily movement: Don is trying to wait patiently, Jane is trying to express disgust by not shaking Sandra's hand. There are also overt actions, which are displayed in bodily movements: Fred's act of climbing the tree is displayed in his body climbing the tree, Samir's act of giving to charity is displayed in her writing the cheque, *etcetera*. In what follows I focus on overt actions, though the results apply more broadly.

There are several views about the relation between overt actions and bodily movements. Some claim that overt actions are identical with bodily movements: "my raising my arm and my arm rising are one and the same event" (Davidson 2004, 103; cp. Anscombe 1957, 12ff). According to this view, reasons cause actions, where actions are identical with overt bodily movements, though these bodily movements can be given action-theoretical descriptions and conceptually inequivalent physical descriptions (Davidson 2004, 101–102). Others claim that overt actions are constituted by, but are not identical with, bodily movements: "Mary's raising of her arm cannot be identified with the motion of her arm" (Hyman 2015, 56; cp. Baker 2012, 255; Bishop 2010, 79; Schlosser 2009, 79).

There are several arguments in support of this latter nonreductive view of actions. First, actions appear to be multiply realizable: Erin's act of hailing a taxi can be realized by one hand waving, two hands forming a stop motion, whistling loudly, *etcetera* (Schlosser 2009, 79). Second, actions appear to involve an "actish phenomenal quality" (Ginet 1990, 11) or an "experience of acting" (Searle 1983, 87; cp. Shepherd 2017) that mere bodily movements do not involve: Jada feels herself trying to raise her arm when she raises it on purpose, but if a stranger pushes her

arm up to the same degree, she does not feel herself trying to raise her arm. Third, the same bodily movement can occur without the bodily movement being an action: Monique's bodily reflex that accidentally kicks the doctor who is testing her reflexes is a bodily movement, but is not Monique's act of kicking.[10] As Bishop explains, "Whenever someone raises her arm, there is the event of her arm going up. She could not have raised her arm without her arm going up; but her arm could have gone up without her raising it" (Bishop 2010, 77). Fourth, the reductive model of action, especially as espoused by Davidson, faces the so-called quausation problem. According to the quausation problem, events cause and are caused in virtue of their properties (cp. Honderich 1982, 63; Sosa 1984, 277; Kim 1984, 267; Horgan 1989, 51). The slipper, in virtue of its fleeciness, not in virtue of its mauvish colour, causes the foot's warmth. Many likewise take brain events and bodily movements to cause and be caused in virtue of non-identical mental and physical properties (Marras 1998, 447; Horgan 1989, 48ff). The physical properties of brain events cause the physical properties of bodily movements, while the mental properties of brain events cause bodily movements to be actions. This widely endorsed quausal model is consistent with the nonreductive model of action. For these reasons, I will take it that a mere bodily movement p^* is not identical to the bodily movement p^* that has the property m^* of being some specific action.

Here is how this distinction between overt action and mere bodily movement aligns with the previously established principles of irreducibility, supervenience, physical causal completeness, and mental causation. As irreducibility stipulates that mental properties m (i.e., being reasons, tryings or intentions) of brain events are not identical with physical properties p (i.e., being physical processes) of brain events, so irreducibility stipulates that mental properties m^* (i.e., being actions) of bodily movements are not identical with physical properties p^* (being musculoskeletal motions) of bodily movements. As supervenience stipulates that mental properties m of brain events supervene upon physical properties p of brain events, so irreducibility stipulates that mental properties m^* of bodily movements supervene upon physical properties p^* of bodily movements. According to physical causal completeness, the physical properties p of brain events are sufficient causes of the physical properties p^* of bodily movements. Mental-to-physical causation states that mental properties m of brain events cause physical properties p^* of bodily movements. But, according to the reasoning outlined above, this renders the physical properties p^* of bodily movements overdetermined by both physical causes p and mental causes m. Thus, the autonomist suggests that physical properties p of brain events cause the physical properties p^* of bodily movements, and the mental properties m of brain events do not cause the physical properties p^* of bodily movements. Rather, the mental properties m of brain events

[10] Granted Monique's mere bodily reflex may be qualitatively distinct from Monique's act of kicking—Monique's reflex may be more twitchy, spasmatic, and may contain slightly different muscular contractions than Monique's act of kicking. These differing bodily movements may occur because they were caused by slightly different neural processes. It is possible, however, to imagine a neuroscientist stimulating the exact neural processes responsible for causing Monique's act of kicking, leading to a qualitatively the same bodily movement as Monique's act of kicking, though this bodily movement is not Monique's act of kicking, since it was caused by the neuroscientist's stimulation of her brain.

cause the mental properties *m** of bodily movements. For example, Monique's physical processing *p* causes her accidental bodily reflex *p** of kicking the doctor, but since Monique was not trying *m* to kick the doctor, her kicking the doctor did not have the property of being an act *m** of kicking the doctor. Had she tried to kick the doctor, her trying to kick the doctor *m* would cause her leg movement *p** to have the property *m** of being her kicking action.

It is worth exploring the third argument raised above for the distinction between overt action and bodily movement in greater detail. Among other things, action theory seeks to answer the following question: what is the difference between bodily movements and action? As Richard Taylor summarizes: "The 'problem of action' … is essentially that of supplying the difference between mere bodily motions and those that represent acts … that there is a difference is perfectly obvious … the fact that such a motion occurs never entails that he makes it, or that it is his act" (Taylor 1960, 88–89). The distinction between whether Monique's leg rising and kicking the doctor was a mere bodily movement, or whether Monique's leg rising and kicking the doctor was a bodily movement constituting her raising her leg to kick the doctor is straightforward. As Oliver Holmes famously states, "even a dog distinguishes between being stumbled over and being kicked" (Holmes 1963, 7). Moreover, this exact distinction is frequently of great moral, legal and pragmatic import—as presumably it would be to Monique's doctor.

In an attempt to answer this question, the standard conception of action in action theory states that a necessary condition for the occurrence of an action is an agent's intention to perform that action. Few doubt the standard conception of action. Here are some famous articulations of the view:

> Tripping over a rug is normally not an action; but it is if done intentionally. Perhaps, then, being intentional is the relevant distinguishing mark (Davidson 1980, 44).

> Not every bodily movement counts as an action—not even those of normal adult human beings—since there are reflex movements, the activities of those who walk in their sleep, and the behaviour of those under hypnosis … it appears as though an action were a bodily movement of a special sort and that we need only to specify the distinctive features of bodily movements that count as actions in order to elucidate the concept of action. We are inclined, accordingly, to look for certain psychological factors in order to mark off bodily movements that count as actions from all those that do not (Melden 1964, 58).

Not only is the standard conception of action not under substantial dispute, but there is also a majority of action theorists who endorse the following standard story of action, or standard theory of action:

> *Causal Theory of Action*: A necessary condition for the occurrence of an action is an agent's reasons causing the action.[11]

The causal theory of action, while under some dispute, has been the dominant model in action theory since Donald Davidson's seminal work *Actions, Reasons and Causes* (1963). Here are some emblematic quotes:

> Suppose an agent acts in some way. What makes it the case that he acted, as distinct from his having been involved in some mere happening or other? … According to the standard story of action that gets told by philosophers, the answer lies in the causal etiology of what happened …We then establish whether the agent acted by seeing whether this bodily movement was caused and rationalized in the right kind of way by some desire the agent had that things be a certain way and a belief he had that something he can just do, namely, move his body in the relevant way, has some suitable chance of making things the way he desired them to be. If so, then that bodily movement is an action; if not, then it is not (Smith 2010, 47).

> So, on this view, an event constitutes an action only if it is caused by rationalizing mental events, and if it does constitute an action, it does so in virtue of being caused by them. The causal history is therefore part of an action's essence or identity. Actions, in other words, are etiological phenomena, just like banknotes, sunburns or Picasso's paintings (Mele 1997, 5).

According to the standard causal theory of action, a necessary condition on an event being an action is that it was caused (in an appropriate way) by mental properties of events such as reasons, intentions, or tryings. Thus, if a bodily movement was not caused by mental properties of brain events, it cannot be an action. In this way, actions are analogous to sunburns: the skin's radiation burn caused by the sun is a sunburn, but an otherwise identical burn, caused by a tanning bed, is simply not a *sunburn*. Or again, while a twenty-dollar bill, created by the American banking institution is a legal tender American twenty-dollar bill, an otherwise identical forgery with all the same etches on the same cotton/linen base, but caused by some counterfeiting cartel, is not a legal tender American twenty-dollar bill. Likewise, Monique's kicking the doctor, caused by her intention to kick the doctor, is an action, but an otherwise identical kicking, caused by Jane's reflex without intention to do so, is not an action.

There are several ways of cashing out the causal theory of action. In the passage above, Mele suggests that the reasons are part of the essence of the action. Support for this view derives from considering intentions that are synchronic with the action, which, since the intention occurs at the same time as the action, allows the possibility that the intention is part of the action. For example, Wanda intends to give a good speech, which persists for the duration of the speech act. There are

[11] For recent book-length defenses of the Causal Theory of Action, consult Bishop (1989), Brand (1984) and Mele (1992). The causal theory of action faces numerous criticisms and issues, but it lies outside the scope of this paper to substantially address these issues. For a recent collection of essays discussing the merits, demerits, and issues surrounding the causal theory of action, see Aguilar and Buckareff (2010).

reasons, however, to think that the intention is not part of the action. First, intentions typically form prior to actions, indicating that intentions are not parts of actions. Even in the Wanda case, she formed the intention to give a good speech some time before she actually began the speech act. Or, more commonly, Joan intends to buy a ticket in the morning, then later in the day she actually does so. Second, intentions can form without the corresponding action occurring. Wanda intends to float during her speech, but the floating act never occurs, again indicating that intentions are not parts of actions. Finally, it is essential to the causal theory of action that reasons are causes of actions. As discussed above, however, causes are temporally and mereologically independent of their effects. Parts, however, are neither temporally nor mereologically independent of their wholes. So, if reasons are causes of actions, then reasons cannot be parts of actions. For these reasons, I prefer not to say that the reason is part of the action, but rather that a necessary condition for an action to occur is that it was caused by reasons.

The standard causal theory of action breathes fresh life into the case for mental-to-mental causation. According to the causal theory of action the mental effect m^* must have some mental cause m. If some mental cause m does not occur, then the mental effect m^* does not occur, even if the physical effect p^* occurs. The causal theory of action, therefore presents a formidable countervailing intuition to Edwards's Dictum. Where Edwards's Dictum insists that p^* is a sufficient determinant of m^* at an instant, rendering m unnecessary as a determining cause of m^*, the causal theory of action insists that m is necessary as a determining cause of m^*. Returning to the sunburn analogy: Edwards's Dictum correctly states that the skin's radiation burn is determined by the DNA damage at that instant, regardless of the fact that ultraviolet radiation caused the burn earlier in the afternoon. But, the skin's radiation burn is not a sunburn if it was not caused by the sun's ultraviolet radiation but was instead caused by a tanning lamp. Similarly, it is true that Jane's musculoskeletal movements determine that Monique's leg rising occurs, regardless of the causal processes leading to the leg rising. But, the leg rising is not Monique's act of raising her leg if it was not caused by her intention to raise her leg, but was instead caused by reflex.

5 Restoring Mental-to-Mental Causation

The argumentation in Sect. 4 shows that the causal theory of action serves as strong motivation for mental-to-mental causation, despite the concern from Edwards's Dictum that m is not a cause of m^*. Matters are not settled, however, as competing intuitions presently obtain. The intuition that p^* is sufficient to determine m^* so m is not necessary as a cause of m^* is supported by Edwards's Dictum, and is also consistent with Supervenience. On the other hand, the intuition that m is necessary as a cause of m^* is supported by the Causal Theory of Action. In this Section I resolve this dilemma by concluding that, while m is necessary as a cause for m^*, supervenience, though not the entirety of Edwards's Dictum, is nevertheless true as well.

Perhaps the clash between the Causal Theory of Action and Edwards's Dictum can be resolved by refuting Edwards's Dictum, thereby paving the way for

the Causal Theory of Action to support mental-to-mental causation. Indeed, Edwards's Dictum has not received favourable press (Campbell 2015, 57ff; Walter 2008, 676; Burge 2007, 372–373), for several reasons. First, Edwards's Dictum entails that the only causal relations that exist are causal relations at the micro-physical level between p and p^*, since they are the only causal processes lacking subvenience bases to jeopardize their causal efficacy. Hence all intralevel macro-physical causation, not just mental-to-mental causation, are excluded—a dubitable result. Second, as discussed above, the horizontal relation is causal while the vertical relation is a non-causal determinative relation, which strains the view that the vertical relation excludes the horizontal relation. The present shape of the statue of David is completely determined by the microstructural properties of its marble, but it is not intuitive that this fact would exclude Michelangelo from having carved the statue hundreds of years ago. For these reasons, and a third reason provided below, Edwards' Dictum is problematic, strengthening the case that mental-to-mental causation is established by appeal to the Causal Theory of Action.

These considerations suffice to show that Edwards's Dictum, which posits the definitive exclusion of m from causing m^*, is problematic. However, recall that Kim generates similar pressure on mental-to-mental causation by appealing to supervenience alone. According to supervenience, p^* is a sufficient determinant of m^*, which implies that m is unnecessary as a cause of m^*, which jeopardizes mental-to-mental causation. The clash, then, is between the Causal Theory of Action, which states that m is necessary as a cause of m^*, and supervenience itself, which implies that m is not needed as a cause of m^*. It is not so easy to dismiss supervenience, as supervenience is partly constitutive of nonreductive physicalism, so the failure of supervenience may entail the failure of nonreductive physicalism. Here is how Jeff Engelhardt frames the worry: "…any view that says there are mental phenomena that do not metaphysically depend on physical phenomena is not a physicalist view… In competitions between mental causes like m and later physical determiners like p^*, either m or p^* loses. If m loses, there is no mental causation. If p^* loses, physicalism is false" (Engelhardt 2015, 226). Presumably physicalism doesn't fail, so supervenience doesn't fail, so the difficulty in securing the mental-to-mental causation posited by the causal theory of action persists.

Fortunately, it is possible to endorse both supervenience and the causal theory of action. Supervenience, as it pertains to m^*, amounts to the truth of the following two counterfactuals:

(5) $p^* \mathbin{\square\!\!\rightarrow} m^*$
(6) $\sim\cup p^* \mathbin{\square\!\!\rightarrow} \sim m^*$

Counterfactual (5) states that if p^* had occurred, m^* would have occurred, which is consistent with the definition of supervenience above stating that 'p necessitates m', or p determines m. Counterfactual (6) states that if none of m's physical realizers occurred, m^* would not have occurred', which is consistent with the definition of supervenience above stating that m occurs 'in virtue of the fact' that

'a physical base p' occurs, or m is dependent upon some p.[12] Counterfactual (6) is true since the nearest worlds where none of m's physical realizers occurs are worlds where no m^* occurs either. In all the nearby worlds where there are no physical realizers of Jane's hailing a cab, Jane's act of hailing a cab does not occur—no disembodied actions occur. Counterfactual (5) is true since the nearest worlds where p^* occurs are worlds where p^* subvenes m^*, and where p occurs to cause p^* and to determine that m occurs. These worlds are exceedingly close to, and in fact include, the actual world. In all the nearest worlds where the physical realizer p^* of Jane's hailing a cab occur, the physical properties of Jane's brain event determined Jane's intention to hail a cab to occur, and caused her bodily movement to occur.

It is possible to maintain the truth of these two counterfactuals while simultaneously endorsing the two counterfactuals that supported mental-to-mental causation. Here they are again:

(1) $m \mathbin{\Box\!\!\rightarrow} m^*$
(2) $\sim m \mathbin{\Box\!\!\rightarrow} \sim m^*$

Counterfactual (1) is uncontroversial: the nearest m-worlds are the worlds where p determines m while p causes p^*, which determines m^* [by (5)], so the nearest m-worlds are m^*-worlds, consistently with the truth of (5). Establishing (2), over and against the truth of (5) is more difficult. Establishing the truth of these two counterfactuals amounts to establishing that the presence of m^* is determined by the presence of p^* [from (5)], yet at the same time, m^* will be absent if m is absent. How can m^* be dependent upon only p^*, while at the same time counterfactually dependent upon m? Fortunately, the appearance of conflict is illusory. Counterfactual (5) is true since all the nearest worlds where p^* occurs are worlds where m^* occurs—these worlds require virtually no changes, so they are exceptionally close. Counterfactual (2) is true since the nearest worlds where m fails to occur are, by supervenience, worlds where no p occurs, so, by completeness, no p^* occurs, so, by supervenience, no m^* occurs. The result: supervenience is true, so p^* determines m^*, but this is compatible with the causal theory of action, so m is also necessary as a cause of m^*.

One can object that m^* now appears overdetermined, as m^* has both a cause m and is fully determined by p^*. According to standard accounts of overdetermination, m^* is overdetermined by m and p^* if the following two counterfactuals are true:

(7) $(\sim m \mathbin{\&} p^*) \mathbin{\Box\!\!\rightarrow} m^*$
(8) $(m \mathbin{\&} \sim p^*) \mathbin{\Box\!\!\rightarrow} m^*$

[12] It is possible to argue that Supervenience is the view that both $p^* \mathbin{\Box\!\!\rightarrow} m^*$ and $m^* \mathbin{\Box\!\!\rightarrow} \cup p^*$ are true. While Supervenience may entail the truth of these counterfactuals, this latter counterfactual does not fully capture the supervenience claim that m^* is dependent upon, because determined by, some p^*. Hence, I substitute this latter counterfactual for counterfactual (6). Nothing of substance below hinges on this substitution.

Counterfactual (8) is true, since the nearest worlds where *m* occurs but *p* fails to occur are worlds where some other physical realizer subvenes *m**. At first glance, counterfactual (7) seems true as well. From (5), supervenience states that the nearest worlds where *p** occurs are worlds where *m** occurs. Counterfactual (7) adverts to a *p** world, so *m** should occur, satisfying (7). This would not only establish the overdetermination result, but may render the causal theory of action false as well, as the overt action *m** can occur without *m* being present to cause it.

Fortunately, the causal theory of action can be sustained while simultaneously avoiding overdetermination. The truth of the causal theory of action provides reason to conclude that counterfactual (7) is false. From (2), the causal theory of action states that the nearest worlds where *m* is absent are also worlds where *m** is absent. Counterfactual (7) adverts to a world where *m* is absent, so *m** should not occur, so (7) is false. This result not only secures the causal theory of action, but also shows that *m** is not overdetermined.

Unfortunately, this result seems to lead to the failure of supervenience. The falsity of (7) shows that there are worlds where *p** occurs, but *m** does not occur, which violates counterfactual (5), which is a necessary condition for establishing supervenience. There are four possible avenues of response. First, it is possible to endorse both the truth of (5) and the falsity of (7). This happens in the case where the (*p** & *m**)-worlds are closer than the (~*m* & *p** & ~*m**)-worlds. It seems clear that the (*p** & *m**)-worlds, where supervenience is not violated, are closer than the (~*m* & *p** & ~*m**)-worlds, which would be distant zombie worlds where supervenience fails. So, (5) is true by virtue of the nearest *p**-worlds being *m**-worlds. At the same time, (7) is false, by virtue of the fact that these farther (~*m* & *p**)-worlds are ~*m**-worlds (since supervenience has failed in these worlds). Since (7) is false, *m** is not overdetermined by both *m* and *p**. The falsity of (7) also satisfies the intuitions grounding the causal theory of action, according to which the occurrence of the same bodily movements *p** without the intent to act *m*, will not yield the action *m**. Thus, Monique's leg kicking, without being intended, is not an action, but a mere bodily movement.

Here is an objection: this solution assumes that (*p** & ~*m**)-worlds, while nomologically impossible, are metaphysically possible. In other words, this solution assumes that the supervenience relation, while nomologically necessary, is not metaphysically necessary. While autonomists may be satisfied with this result, many nonreductive physicalists reject the view that (*p** & ~*m**)-worlds are metaphysically possible—in fact, some consider this view partly constitutive of nonreductive physicalism (Loewer 2015, 61). If this is true, there are no (*p** & ~*m**)-worlds, so all the *p**-worlds are *m**-worlds, so (5) cannot be true while (7) is false.

There are still three remaining avenues of response available, each of which is consistent with a nonreductive physicalism that endorses a metaphysically necessary supervenience relation. One can accept the currently flourishing prospect that impossible worlds exist and can be ordered (Bjerring 2014; Brogaard and Salerno 2013). According to this view, one can reason non-trivially despite impossible antecedents. To borrow a common example, the impossible world in which Hobbes squared the circle and the mathematicians were thrilled is closer to our world than the impossible world in which Hobbes squared the circle and the

sick children in Afghanistan were thrilled. The former case is impossible but reasonable, where the latter case is both impossible and unreasonable. With this in mind, it is clear that the (p^* & m^*)-worlds, where supervenience is not violated, are closer than the (~m & p^* & ~m^*)-worlds, which would be extremely remote logically impossible worlds where supervenience fails. So, (5) is true by virtue of the nearest p^*-worlds being m^*-worlds. At the same time, (7) is counterpossibly unreasonable, by virtue of the fact that these impossible (~m & p^*)-worlds are ~m^*-worlds (since supervenience has failed in these worlds). Since (7) is false, m^* is not overdetermined by both m and p^*. The counterpossible unreasonableness of (7) also satisfies the intuitions grounding the causal theory of action, according to which the occurrence of the same bodily movements p^* without the intent to act m, will not yield the action m^*.

For nonreductive physicalists averse to impossible worlds, it is also possible to conclude that (5) is true while (7) is vacuous. That is, supervenience is metaphysically necessary, so all the worlds where p^* occur are worlds where m^* occurs, so (5) is true. At the same time, supervenience also stipulates that (6) is true, so all the worlds where m fails to occur are worlds where no p occurs. But from completeness, all the worlds where no p occurs are also worlds where no p^* occurs, so there do not exist any possible (~m and p^*)-worlds, rendering (7) vacuous. The result: since there are no (~m & p^* & m^*)-worlds, all possible m^*-worlds are both some p^*-worlds and m-worlds, which is consistent with both the causal theory of action and supervenience.

Finally, it is possible to take the falsity of (7) to show that (5) is false. In other words, it is possible to claim that the necessity of m as a cause of m^* (from the causal theory of action) does falsify the claim that p^* is a sufficient determinant of m^* (from supervenience) (Schlosser 2009, 87). It appears drastic to doubt supervenience, but there are intuitions suggesting that supervenience states that the subvening base is only the sufficient determinant of the intrinsic macroproperties of the object, not the extrinsic macroproperties of the object. To return to prior examples: the subvening DNA damage at t does determine the skin's intrinsic macroproperties at t such as its redness and its burntness, but the DNA damage does not determine that the skin is sunburnt—how would local microproperties affect whether the distant sun or a tanning bed caused the burn? Or, the tiny etches and scratches in the bill at t does determine the bill has the shape of a 20 on it, and the shape of Andrew Jackson on it, but these etches and scratches do not determine that the bill is legal tender created by the federal reserve rather than a counterfeiting cartel. Similar intuitions preside in the realm of mental causation. Davidson's swampman is a spontaneously generated being (by a bolt of lighting striking a swamp) who is microstructurally hence macrostructurally the same as the real Davidson. But, the swampman "can't recognize anything, nor have any thoughts" (Davidson 1987, 444), since recognition and thinking is "identified in part by their causal relations to events and objects outside the subject" (Davidson 1987, 444). Karen Bennett likewise takes the combination of content externalism with the view that mental states have their contents essentially to yield the view that the subvenience base p does not by itself determine m—since m's content is essentially connected with extrinsic factors (Bennett 2003, 484). Similarly, while Monique's musculoskeletal movements determine the

velocity and direction of her leg rising, whether her leg rising p^* is her act m^* of raising her leg depends on whether it was caused by Monique intention m to raise her leg.

Here is another objection to the argumentation laid out in Sect. 4. Namely, the causal theory of action salvaged mental-to-mental causation from supervenience based exclusion pressures, but mental-to-mental causation may still face upward causation based exclusion pressures. That is, while m causes m^*, despite the fact that p^* determines m^*, it is possible that m fails to cause m^* because p causes m^* instead. This objection presumes the principle of Upward Causation: "If property A causes property B, then … A causes any supervenient property of B instantiated on this occasion" (Zhong 2014, 347; cp. Engelhardt 2015, 211; Pernu 2016, 1042). By completeness, p is a sufficient cause of p^*, and by supervenience p^* completely determines m^*, so, by upward causation, p is a sufficient cause of m^*. So, by exclusion, m is excluded from causing m^*, since p is a sufficient cause of m^*. Mental-to-mental causation is jeopardized anew.

A variety of responses are available to this upward causation problem. First, by supervenience, p^* determines m^*, which is a vertical determinative relation. The relation from p to m^*, however, is defined as a horizontal causal relation. If exclusion pressure arises between vertical determinative relations and horizontal causal relations, Edwards's Dictum states that vertical determinative relations exclude horizontal causal relations. So, according to Edwards's Dictum, p^*, as a sufficient vertical determinant of m^*, would exclude the purported cause p of m^*. So, on Edwards's Dictum, p is excluded from causing m^*, so p is not even a cause of m^*. Surely p cannot exclude m from causing m^* when p is not even a cause of m^*. If there is exclusion pressure with respect to m^*, it is between p^* and m, which is the subject matter of this paper.

This result provides a third problem facing Edwards's Dictum. Namely, Edwards's Dictum entails that microphysical states cannot cause any macrophysical effects, which is an unsavory result. All Edwards's Dictum allows is that microphysical states cause microphysical effects, which themselves determine macrophysical effects.[13] Even if one abandons Edwards's Dictum at this point, the upward causation issue may re-appear. Namely, if m is a cause of m^*, and p is a sufficient cause of m^*, straightforward causal exclusion pressure obtains between m and p. Notice, however, that the upward causal relation between p and m^* is established via establishing that p causes p^* to determine m^*. The relation from p^* to m^* is a non-causal process, so a causal process from p to m^* cannot be established by appealing to a causal chain from p to p^* to m^*, or by appealing to the transitivity of causality (Thomasson 1998, 189–190; Engelhardt 2017, 34–35). So the suggestion that m cannot cause m^* because p causes m^* fails, since it is not established that p causes m^*. Finally, even if these difficulties can be surmounted, Pernu and Zhong have both articulated how mental-to-mental causation can be preserved over and against the

[13] Kim sidesteps these issues by abandoning Irreducibility, which renders moot the discussion on the viability of Upward Causation and Downward Causation, and allows for only horizontal causation.

threat of physical-to-mental causation. Namely, the following two counterfactuals must be true in order to ensure the upward causation of p to m*:

(9) $p \mathrel{\square\!\!\rightarrow} m^*$
(10) $\sim p \mathrel{\square\!\!\rightarrow} \sim m^*$.

Counterfactual (9) is true, since the nearest p-worlds are worlds where both completeness and supervenience hold, so m* occurs. However, (10) may fail, since the nearest $\sim p$-worlds may be worlds where p is replaced by p_2, so p* is replaced by p_2*, which is a slightly different realizer of the same m that occurs (Engelhardt 2015, 224–225; Pernu 2016, 1042–1043; Zhong 2014, 355–356). In this case, (10) is false, so the upward causation from p to m* does not occur, so mental-to-mental causation is not excluded by p causing m*.

In summary, recent attempts by Zhong and Pernu to solve the causal exclusion problem by appealing to an autonomy solution fail to address the problem of Edwards's Dictum. The problem of Edwards's Dictum, however, can be overcome by appealing to central tenets of action theory, and demonstrating how these tenets of action theory can be maintained while also supporting supervenience. This re-establishes the case for mental-to-mental causation, thereby re-invigorating the autonomist solution to the causal exclusion problem.

Acknowledgements I would like to thank the anonymous referees at this journal for their valuable comments and suggestions that greatly improved this paper. I would also like to thank the audience members and commentators at the 2018 *Western Canadian Philosophical Association* Meeting, the 2019 *Canadian Philosophical Association* Meeting, and the 2019 *Central American Philosophical Association* Meeting.

References

Aguilar, J., & Buckareff, A. (2010). *Causing human actions*. Cambridge: MIT Press.
Anscombe, E. (1957). *Intention*. Cambridge: Harvard University Press.
Baker, L. (2012). What we do. In J. Bransen & S. Cuypers (Eds.), *Human action, deliberation and causation* (pp. 249–270). Berlin: Springer.
Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. *Noûs, 37,* 471–497.
Bishop, J. (1989). *Natural agency*. Cambridge: Cambridge University Press.
Bishop, J. (2010). Skepticism about natural agency and the causal theory of action. In J. Aguilar & A. Buckareff (Eds.), *Causing human actions* (pp. 69–84). Cambridge: MIT Press.
Bjerring, J. (2014). On Counterpossibles. *Philosophical Studies*, *168*, 327–353.
Bogardus, T. (2013). Undefeated Dualism. *Philosophical Studies*, *165*(2), 445–466.
Brand, M. (1984). *Intenting and acting*. Cambridge: MIT Press.
Brogaard, B., & Salerno, J. (2013). Remarks on Counterpossibles. *Synthese*, *190*(4), 639–660.
Burge, T. (1993). Mind-body causation and explanatory practice. In J. Heil & A. Mele (Eds.), *Mental causation* (pp. 99–117). Oxford: Clarendon.
Burge, T. (2007). *Foundations of mind: Philosophical essays* (Vol. 2). Oxford: Oxford University Press.
Campbell, N. (2015). Does same level causation entail downward causation. *Abstracta, 8*(2), 53–66.
Cavedon-Taylor, D. (2008). Still epiphenominal qualia. *Philosophia, 37*(1), 105–107.
Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.
Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22,* 547–563.
Crisp, T., & Warfield, T. (2001). Kim's master argument. *Noûs, 35,* 304–316.

Davidson, D. (1963). Actions, reasons and causes. *Journal of Philosophy, 60,* 685–700.
Davidson, D. (1980). *"Agency", actions and events*. Oxford: Clarendon Press.
Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441–458.
Davidson, D. (2004). Problems in the explanation of action. In M. Cavell (Ed.), *Problems of rationality* (pp. 101–116). Oxford: Clarendon Press.
Engelhardt, J. (2015). What is the exclusion problem? *Pacific Philosophical Quarterly, 96,* 205–232.
Engelhardt, J. (2017). Mental causation is not just downward causation. *Ratio, 30*(1), 31–46.
Fodor, J. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese, 28,* 77–115.
Fodor, J. (1989). Making mind matter more. *Philosophical Topics, 17*(1), 59–79.
Gibbons, J. (2006). Mental causation without downward causation. *Philosophical Review, 115,* 79–103.
Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.
Harbecke, J. (2014). Counterfactual causation and mental causation. *Philosophia, 42*(2), 363–385.
Holmes, O. (1963). *The common law*. Boston: Little, Brown.
Honderich, T. (1982). The argument for anomalous monism. *Analysis, 42,* 59–64.
Horgan, T. (1989). Mental quausation. *Philosophical Perspectives, 3,* 47–76.
Hume, D. (1888). *A treatise of human nature*, L. Selby Bigge (ed.), Oxford: Oxford University Press.
Hyman, J. (2015). *Action, knowledge and will*. Oxford: Oxford University Press.
Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly, 32*(127), 127–136.
Jacob, P. (2002). Some problems for reductive physicalism. *Philosophy and Phenomenological Research, 65,* 648–654.
Kim, J. (1984). Epiphenomenal and supervenient causation. *Midwest Studies in Philosophy, 9,* 257–270.
Kim, J. (1995). Mental causation in Searle's biological naturalism. *Philosophy and Phenomenological Research, 55,* 189–194.
Kim, J. (1998). *Mind in a physical world*. Cambridge: MIT Press.
Kim, J. (1999). Making sense of emergence. *Philosophical Studies, 95,* 3–36.
Kim, J. (2005). *Physicalism, or something near enough*. Princeton: Princeton University Press.
Kim, J. (2009). Mental causation. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *Oxford handbook of philosophy of mind* (pp. 29–52). Oxford: Oxford University Press.
Lewis, D. (1973). Causality. *The Journal of Philosophy, 70,* 556–567.
Lewis, D. (1986). *Philosophical papers* (Vol. II). Oxford: Oxford University Press.
List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106*(9), 475–502.
Loewer, B. (2015). Mental causation: The free lunch. In T. Horgan, M. Sabates, & D. Sosa (Eds.), *Qualia and mental causation in a physical world* (pp. 40–63). Cambridge: Cambridge University Press.
Lowe, E. (1993). The causal autonomy of the mental. *Mind, 102,* 629–644.
Lowe, E. (2006). Non-Cartesian substance dualism and the problem of mental causation. *Erkenntnis, 65*(1), 5–23.
Marras, A. (1998). Kim's principle of explanatory exclusion. *Australasian Journal of Philosophy, 76,* 439–451.
Marras, A. (2007). Kim's supervenience argument and nonreductive physicalism. *Erkenntnis, 66,* 305–327.
McDonnell, N. (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies, 174*(6), 1459–1474.
McGinn, C. (1989). *Mental content*. Oxford: Blackwell.
Melden, A. (1964). Action. In D. Gustafson (Ed.), *Essays in philosophical psychology*. New York: Doubleday.
Mele, A. (1992). *Springs of action*. Oxford: Oxford University Press.
Mele, A. (1997). Agency and mental action. *Philosophical Perspectives, 11*(S), 231–249.
Melnyk, A. (2003). *Physicalist manifesto*. Cambridge: Cambridge University Press.
Montero, B. (2003). Varieties of causal closure. In S. Walter & H. Hackmann (Eds.), *Physicalism and mental causation* (pp. 173–187). Imprint Academic: Exeter.
Moore, D. (2013). Counterfactuals, autonomy and downward causation. *Philosophia, 41,* 831–839.
Nagasawa, Y. (2010). The knowledge argument and epiphenomenalism. *Erkenntnis, 72,* 37–56.

Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents* (pp. 3–36). Cambridge: Cambridge University Press.
Pernu, T. (2016). Causal exclusion and downward counterfactuals. *Erkenntnis, 81*(5), 1031–1049.
Putnam, H. (1967). Psychological predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
Schlosser, M. (2009). Nonreductive physicalism, mental causation and the nature of actions. In Alexander Hieke & Hannes Leitgeb (Eds.), *Reduction: Between the mind and the brain* (pp. 73–90). Frankfurt: Ontos Verlag.
Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.
Shepherd, J. (2017). The experience of acting and the structure of consciousness. *Journal of Philosophy, 114*(8), 422–448.
Silberstein, M. (2001). Converging on emergence. *Journal of Consciousness Studies, 8,* 61–98.
Smith, M. (2010). The standard story of action: An exchange. In J. Aguilar & A. Buckareff (Eds.), *Causing human actions* (pp. 45–56). Cambridge: MIT Press.
Sosa, E. (1984). Mind–body interaction and supervenient causation. *Midwest Studies in Philosophy, 9,* 271–281.
Taylor, R. (1960). *Action and purpose*. Englewood Cliffs: Prentice Hall.
Thomasson, A. (1998). A nonreductivist solution to mental causation. *Philosophical Studies, 89,* 181–195.
Tuomela, R. (1998). A defense of mental causation. *Philosophical Studies, 90*(1), 1–34.
Walter, S. (2008). The supervenience argument, overdetermination, and causal drainage. *Philosophical Psychology, 21,* 673–696.
Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.
Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation*. Oxford: Oxford University Press.
Zhong, L. (2011). Can counterfactuals solve the exclusion problem? *Philosophy and Phenomenological Research, 83*(1), 129–147.
Zhong, L. (2014). Sophisticated exclusion and sophisticated causation. *Journal of Philosophy, 111*(7), 341–360.
Zhong, L. (2019). Taking emergentism seriously. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2019.1589547.