

10-page Summary of “Mental Causation and Rational Agency”

I. Introduction

In this dissertation, I am working at the intersection of the philosophy of mind and the philosophy of action. In particular, I am concerned here with a worry about the very possibility of our rational agency, of our capacity to act for reasons, that has received increasing attention in recent years. This worry arises when the contemporary incarnation of a classic problem in the philosophy of mind—the so-called “problem of mental causation”—is combined with a view of the nature of our intentional actions that has become orthodoxy within the philosophy of action since, and largely because of, Donald Davidson’s agenda-setting work in the 1960s and ‘70s. According to the orthodox view, which I call “Metaphysical Causalism,” what it is for an agent’s physical movements to be an intentional action, a manifestation of her capacity to act for reasons, is for them to have a certain causal provenance in her mind. But, the worry goes, if we cannot understand how the mind could have physical effects—which is the threat posed by the contemporary form of the problem of mental causation—then we apparently cannot understand how there could be rational agency. And so, it has seemed to many philosophers that, without a solution to the problem of mental causation, we cannot make sense of rational agency.

I argue for three main theses in this dissertation. The first is that, even if we were to accept the orthodox view of rational agency, it is not clear that there is in fact the difficulty with understanding the mind as having physical effects that the contemporary literature takes there to be. In particular, I argue, the idea that mental causes overdetermine their physical effects is not the unattractive response to the problem that the literature has largely assumed it to be, because the literature has generally failed to distinguish between importantly different senses in which effects may be “overdetermined” by their causes. I use biological and other non-mental examples to suggest that the kind of overdetermination that mental-to-physical causation is a case of is both innocuous and widespread in the natural world.

The deeper issue, however—and this is the second main thesis I argue for—is that the above problematic rests on a mistaken view of the role of causation in rational agency. An agent’s intentional actions can, and no doubt do, have causal antecedents—even mental causal antecedents—but that fact is not explanatory of their being intentional actions. The status of our intentional actions as manifestations of rational agency is not an extrinsic, non-essential feature that is bestowed upon certain

bodily movements by a causal transaction with the agent's mind, as Metaphysical Causalism would have it. Finally, I argue for an alternative account of rational agency that is inspired by the seminal work of Elizabeth Anscombe in *Intention* (1953), and on which the rational mindedness characteristic of our intentional actions is, rather, an intrinsic, essential feature of them. In particular, I argue, a proper appreciation of the sort of explanatory work that is done in those explanations of an agent intentional action that cite her reasons for Φ 'ing ("agent's-reasons explanations") reveals that intentional actions manifest rational mindedness not in virtue of their causal provenance, but in the rationally-ordered, means-end structure that is essential to their being the intentional actions they are. This structure is revealed in correct answers to the question "Why are you Φ -ing?" that asks for the agent's reasons for Φ -ing and that seeks, I argue, an explanation of what (more specific) *kind* of intentional action the agent's (more generically specified) intentional action of Φ -ing is, rather than an explanation of the *occurrence* of the event that the intentional action is, i.e., something that would be explained by the occurrence of an antecedent event. Ultimately, on my account, such things as antecedent episodes of rational deliberation about what to do, and subsequent intentional actions of doing what was decided upon, are essentially related not as cause and effect, but as relatively earlier and later phases of a single unitary process of rational action.

If these theses are correct, then the worry about rational agency will have been addressed in two ways: (1) by uncoupling the question of the nature of intentional action from mental-to-physical causation, and so from the problem of mental causation; (2) by providing a novel overdetermination account for those mental-to-physical transactions that there might be. In the remainder of this abstract, I explain in more detail how the worry about the intelligibility of rational agency arises and can come to seem pressing, and I summarize the arguments I present in each chapter.

II. The Problem of Mental Causation and Rational Agency

With the passing of Cartesian substance dualism, and the ascendance of a scientific view of the world as fundamentally physical, the consensus in the contemporary literature is that the primary obstacle to making sense of mental causation is to show how it is consistent with the plausible idea that, at least in principle, any physical effect can be completely accounted for in exclusively physical terms,

without appealing to any causes from outside the physical domain. This conception of the physical domain is thought to pose an obstacle to making sense of mental causation because it appears to form an inconsistent triad when combined with two other widely accepted ideas. The first is the familiar idea that mental phenomena are not part of the physical domain. On one popular way of putting the point, it makes sense to treat all the various phenomena of a person's mental life as properties of her, and these mental properties fail to be identical with, or reducible to, her or anything else's physical properties. The other thought, to complete the inconsistent triad, is the very idea of mental causation and, in particular, the idea of "mental-to-physical" causation, i.e., the idea that mental phenomena can and do have physical effects. The intuitive tension, then, is this: on the one hand, the physical domain appears to be a causally seamless and self-contained realm, with neither need, nor room, for causal influences from other domains; there are no gaps in the chains of physical causes leading to any physical effect, and the constituents of the relevant causal chains are jointly sufficient to bring about that effect. On the other hand, mental phenomena appear to constitute just such a nonphysical domain, and yet they also nevertheless appear to have physical effects. And thus we can see the contemporary guise of a classic philosophical conundrum: the problem of mental causation.

Now, suppose we accept that there is a problem with making sense of mental causation. Why should we care? That is, what motivates the urgency that is so often felt around the problem of mental causation? What larger concerns are supposed to be at stake in our being able to make sense of mental phenomena as causes of physical effects? In the literature on mental causation, the motivation that is perhaps most often offered to underwrite the urgency of the problem of mental causation is a worry about the possibility of making sense of our status as agents—in particular, as *rational* agents, i.e., as creatures who act for the sake of reasons—if we cannot make sense of mental phenomena as having physical effects. According to Jaegwon Kim, for example, our recognition of a causal relationship between our mental lives and the physical movements of our bodies when we act is what grounds our practice of assigning moral responsibility, praise, and blame, as well as our understanding of ourselves as persons who act for reasons at all. If this is right, then the possibility of making sense of ourselves as rational agents depends on our being able to make sense of mental causation.

The picture of intentional actions that Kim and other philosophers of mind take as uncontroversial, when describing why a solution to the problem of mental causation is so urgent, is precisely the view that many philosophers of action see as the only viable answer to a question that they take to be definitive of their field of philosophical inquiry. As the question is often put, adapting a passage from Wittgenstein's *Philosophical Investigations*: What is left over if the fact that my arm went up is subtracted from the fact that I raised it? As philosophers of action standardly conceive of their work, to answer this question is to explain the nature of, or what it is to be, an intentional action, as well as how intentional actions differ from other things agents do that are not intentional actions. According to the view that has become orthodoxy within philosophy of action, what it is to be an intentional action is to be a bodily movement that was caused in a certain way by the agent's mind. Those bodily movements that have the appropriate mental causal provenance count as intentional actions; those that do not are not intentional actions. Call this view "Metaphysical Causalism" about intentional actions.

The foregoing considerations make it possible for us to now reconstruct a simple, three-step argument against the possibility of rational agency that I will call the "Organizing Argument" because the work of this dissertation is organized around it:

- (1) Rational agency constitutively involves mental causation; specifically, an agent's mental phenomena cause the bodily movements that are her intentional actions.
- (2) Bodily movements are physical phenomena.
- (3) Mental phenomena cannot have physical phenomena as effects.
- (4) Therefore, there can be no such thing as rational agency.

In fact, however, I think our self-understanding as rational agents who act for the sake of reasons is not at all in danger, at least from the considerations I have discussed here. This is because I think—and it is the negative project of this dissertation to argue—that premises (1) and (3) in the Organizing Argument are false. My purpose in doing so is at each stage to displace key background premises that hold the Metaphysical Causalist picture in place and thereby to progressively recast our philosophical understanding of rational agency. The positive project of this dissertation is to motivate and explicate an alternative conception of rational agency whereon our intentional actions manifest our

rational mindedness in the rationally-ordered means-end structure that is essential to their being the intentional actions they are. My view thus defuses the direct threat to the intelligibility of rational agency that is posed by the putative need to explain how the mind can cause bodily movements, because acting for a reason is not constituted by such a causal transaction. It also accounts for how one's mind can have physical effects by overdetermining them alongside their physical causes.

III. Chapter-by-Chapter Summary of the Argument of the Dissertation

In Chapter One, I provide an overview of the argument of this dissertation. I begin the substantive work of the dissertation in Chapter Two, where I argue that Premise (3) of the Organizing Argument is false. In particular, I argue, even if it were true that rational agency constitutively involves mental causation; and even if it were true that the effects of mental causes in cases of rational agency are physical phenomena; mental causes can still have physical effects because they can systematically overdetermine those effects, alongside those effects' sufficient physical causes. The idea that mental phenomena systematically overdetermine their physical effects remains largely undeveloped in the literature. The reason for this state of affairs is not far to find: the idea of systematic mental-to-physical overdetermination is widely felt (though seldom argued) to be unattractive, implausible, and even *ad hoc*. I argue that such sentiments are mistaken, however, and that they are rooted in an inappropriate and unmotivated model of what mental-to-physical overdetermination would have to involve. Specifically, I argue that the literature has generally failed to distinguish between importantly different senses in which effects may be "overdetermined" by their causes, and so it has failed to appreciate that the overdetermination of physical effects by their mental causes need not—and, indeed, ought not—be understood on the model of a classic "firing squad" case. Rather, I argue, mental-to-physical causation is in fact a (special) case of a quite different variety of causal overdetermination, a variety that obtains when the set of causes of a given effect merely includes more than is nomologically necessary to bring about the effect, but without that fact implying that the relevant effect has more than one individually sufficient cause. I call this sort of overdetermination "super-sufficiency." Using examples from biology and other non-mental cases, I show that super-sufficiency is both innocuous and widespread in the natural world, even apart from mental-to-physical causation.

In Chapter Three, I begin my critical assessment of Premise (1) by rebutting the so-called “Master Argument” that is widely regarded as the linchpin of Metaphysical Causalism. The argument in question is the one Donald Davidson presents in “Actions, Reasons, and Causes” (1963) for the view that a reason R that an agent S (takes it she) has to Φ is S 's reason for Φ -ing only if R causes S 's Φ -ing. Proponents of Metaphysical Causalism generally take Davidson's Master Argument to conclusively establish that it is necessary that an intentional action have a mental cause. I argue, however, that this argument involves a mistaken view of the explanatory work we are doing when we explain an agent's intentional action of Φ -ing by citing her reasons for Φ -ing. In particular, Metaphysical Causalism mistakenly assumes that what we are in general doing, when we explain S 's intentional action of Φ -ing by giving her reasons for Φ -ing, is explaining the *occurrence* of an event that is the intentional action in question, i.e., something that would be explained by the occurrence of an antecedent event.

Rather, I argue, what we are doing in agent's-reasons explanations—at least in a wide variety of central cases—is explaining the (more specific) *kind* of intentional action that S 's (more generically specified) intentional action of Φ -ing is, explanatory work that is paradigmatically done by specifying the intentional undertaking of which S 's Φ -ing is a constitutive part, e.g., S 's Ψ -ing. Thus, for example, when we receive the (correct) answer, “To buy milk” in response to the agent's-reasons-requesting sense of the question “Why is S going to the store?”, we are not in the first instance (indirectly) finding out what antecedent event explains the occurrence of the event that is S 's intentional action of going to the store. Rather, we are finding out that S 's going to the store is a constitutive part of her buying milk, and thus that her going to the store is a going-to-the-store-to-buy-milk, as opposed, say, to a going-to-the-store-to-kill-the-clerk. Once this distinction between explaining *kind* and explaining *occurrence* is clearly in view, there is little attractiveness to the idea that agent's-reasons explanations are causal explanations, and we can see that answers like, “To buy milk,” are perfectly in order to do the sort of explanatory work that the relevant sense of the question “Why is S going to the store?” asks to be done, without being underwritten by a causal transaction between an agent's mind and her bodily movements.

Having thus argued that Davidson's Master Argument does not motivate Metaphysical Causalism in the way that many philosophers of action have thought it does, in Chapters Four and Five I

advance my case further against Metaphysical Causalism, presenting a two-part argument not just that Metaphysical Causalism is not well-motivated, but that it in fact looks to be hopeless, given its aim of providing a substantive, non-circular account of what it is to be an intentional action. In Chapter Four, I take up the mental phenomena that Metaphysical Causalism proposes as causes, and I argue that both intentions to Φ and reasons that agents have for Φ -ing are essentially intentional-action-directed: when an agent S intends to Φ , she intends an intentional action, and when what an agent does is at issue, S has reasons only for her intentional actions. If the former is right, then it generates an unacceptable circularity for that variant of Metaphysical Causalism that would analyze what it is for S 's Φ -ing to be an intentional action in terms of an appropriate causal relation to S 's intention to Φ . And if the latter is right, then the same goes for that variant of Metaphysical Causalism that would analyze what it is for S 's Φ -ing to be an intentional action in terms of an appropriate causal relation to certain of S 's beliefs and desires, since (according to this variety of Metaphysical Causalism) the relevant beliefs and desires just are S 's reasons for Φ -ing, and thus are essentially intentionally-action-directed. Moreover, if philosophers like Myles Brand (1984), Alfred Mele (1992), and others are correct that a constitutive account of intentional action cannot be provided solely in terms of causal relations to beliefs and desires of the agent, and that such an account must also make essential reference to agent's intentions, then the intention-based circularity is unavoidable here, too.

In Chapter Five, I turn to the issue of how to understand the nature of the relevant effects in Metaphysical Causalism's proposed account, and I introduce the topic by returning to the idea of defining an intentional action as one done for a reason that, as I showed in Chapter Three, plays a crucial role in the argument for Metaphysical Causalism. In the proposed definition, "one" is naturally read as "action", which suggests a first pass at an analysis of what it is to be an intentional action—namely, S 's Φ -ing is an intentional action iff S 's Φ -ing is an action done for a reason—which then awaits only a causal analysis of what it is for an action to be done for a reason in order to complete a causal analysis of intentional action. The problem, however, is that the first-pass biconditional can even begin to look true only if we implicitly read "action" on the right-hand side of the biconditional as "intentional action": in the domain of things we do, the only things it makes sense to say that we do "for reasons," in the

relevant sense of “for the sake of reasons,” are intentional actions. However, if we make that implicit reading explicit—such that the first-pass biconditional reads “*S*’s Φ -ing is an intentional action iff *S*’s Φ -ing is an *intentional action* done for a reason”—it becomes clear that, even if it is true, the first-pass biconditional is not a step in the direction of a substantive, non-circular account of the sort that Metaphysical Causalism purports to provide.

The deep issue here, I argue, is an untenable presupposition of Metaphysical Causalism, viz., that there is such a thing as what I call a “neutral item.” As I use the term, a neutral item would be something that: occurs both in the case of intentional actions, and in the case of unintentional actions; which is neutral, as the kind of occurrence it is, between its being intentional or unintentional; such that its being intentional or not in a given case can be determined (as Metaphysical Causalism holds it is) by whether it has a certain mental causal provenance. Thus, for example, in the context of Wittgenstein’s question that I discussed in Section II and that sets the agenda for much of contemporary philosophy of action, my arm’s rising would be the “neutral item” which occurs both in the case of intentional actions, and in the case of unintentional actions, and which is neutral, as the kind of thing it is, between being intentional and being an unintentional, such that its being an intentional action or not could be determined by whether it has a certain causal provenance in the mind of the agent.

However, I argue, the idea that there are such neutral items is just an illusion induced, in part, by the availability of descriptions of things agents do that are silent as to whether the things described are done intentionally or not. As I have already argued in Chapter Four, however, such apparently neutral descriptions are typically abbreviated descriptions of the form “*S*’s Φ -ing in order to Ψ ,” which latter descriptions make plain that what is being described is an intentional action. In fact, the doings that are intentional actions have, as such, a rationally-ordered, means-end structure: an intentional action, as the kind of doing it is, is done for the sake of another such doing, or is that for the sake of which some other such doing is done.

Or, to put the point in the terms Harry Frankfurt introduces in his seminal 1978 essay “The Problem of Action,” the untenable presupposition in Metaphysical Causalism is the idea that events that are intentional actions, and events that are not intentional actions, “do not differ essentially in

themselves at all”; intentional actions are “inherently indistinguishable” from events that are not intentional actions, such that their being intentional is an extrinsic property (Frankfurt 1978:157). However, I argue, this is false: there are internal, intrinsic facts, facts about intentional actions as such, that distinguish intentional actions from things that are not intentional actions. In particular, what distinguishes intentional actions, as such, from things that are not intentional actions is the rationally-ordered means-end structure that is essential to their being the intentional actions they are. As a matter of the kind of happening they are, intentional actions have the form “ Φ -ing (in order) to Ψ ” or “ Φ -ing by (means of) Ψ -ing”. This internal structure of intentional actions is not shared with events that are not intentional actions. It is, therefore, not true that the events that are intentional actions are “inherently indistinguishable” from events that are not intentional actions.

In the concluding chapter of this dissertation, Chapter Six, I have two interrelated objectives. The first is to clarify the non-reductive conception of intentional action that I am defending in this dissertation as an alternative to Metaphysical Causalism. My second objective—and the means by which I propose to accomplish the first—is to distinguish my view and arguments from those of other recent philosophers of action who have also drawn upon the work of Elizabeth Anscombe to argue against Metaphysical Causalism and in favor of a non-reductive conception of intentional action. In particular, I distinguish my view from that of Douglas Lavin and I critically evaluate two arguments against Metaphysical Causalism that he develops in two recent papers. I argue that these arguments are unnecessary to resist Metaphysical Causalism, fail to establish their conclusions, and would, if accepted, commit us to an implausible picture of intentional action, on which every intentional action consists of an infinite number of subordinate intentional actions that are done for its sake.

With that critical work as background, I sketch a summary picture of my alternative to Metaphysical Causalism that incorporates the various positive points that I have developed in previous chapters. To bring out its distinctive features, I contrast my position with a position that I call “Psychological Causalism” about intentional action, and which might appear attractive to someone who has been convinced by my arguments to abandon Metaphysical Causalism, but who still would like to retain an essential place for causation in our understanding of intentional action, perhaps as a way of

securing an unproblematic place for it in the natural order. I suggest, by contrast, that the relation between, e.g., antecedent episodes of rational deliberation about what to do, and subsequent intentional actions of doing what was decided upon, is like the relation between acorns and the oak trees that they result in (but do not cause), or between the prophase of mitosis and the metaphase of mitosis that it leads to (but does not cause): they are relatively earlier and later stages, respectively, of a unitary process whose principle of unity is not causal, but rationally teleological. Indeed, as I argued in Chapter Four, that this is so is precisely what consideration of agent's-reasons explanations would suggest: agent's-reasons explanations explain intentional actions by revealing the larger intentional undertakings of which they are constitutive parts, and which, in turn, partially determine the kind of intentional action the *explanandum* ultimately is. Going-to-the-store-to-buy-milk is a different kind of intentional action than going-to-the-store-to-kill-the-clerk, and what is explained to us, when we are given a (correct) response to the agent's-reason-requesting sense of the question "Why is *S* going to the store?" is what (more specific) kind of going to the store *S*'s (more generically specified) intentional action of going to the store is.

Once this internal, rationally-ordered structure is made explicit in a (correct) answer to such a Why-question—or to a complementary How-question that asks about *S*'s means to those ends—and so once we have available an unabbreviated description of the relevant intentional action that wears its internally-grounded status as an intentional action on its sleeve, e.g., "*S*'s Φ -ing (in order) to Ψ " or "*S*'s Φ -ing by (means of) Ψ -ing", there is no need to then appeal to causal relations between *S*'s Φ -ing and anything else to make sense of it as an intentional action, as a manifestation of our rational capacity to act for reasons.