AGENT CAUSATION AND REDUCTION

BY

SUN HYUNG RHEE

B.A. HANYANG UNIVERSITY, 1998
M.A. SEOUL NATIONAL UNIVERSITY, 2001

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE DEPARTMENT OF PHILOSOPHY AT BROWN UNIVERSITY

PROVIDENCE, RHODE ISLAND

MAY 2014
© Copyright 2014 by Sun Hyung Rhee
This dissertation by Sun Hyung Rhee is accepted in its present form by the Department of Philosophy as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date: ____________________  ____________________
       Jaegwon Kim, Advisor

Recommendation to the Graduate Council

Date: ____________________  ____________________
       Bernard Reginster, Reader

Date: ____________________  ____________________
       Mary Louise Gill, Reader

Approved by the Graduate Council

Date: ____________________  ____________________
       Peter Weber, Dean of the Graduate School
CULICULUM VITAE

Sun Hyung Rhee received her B.A. in philosophy from Hanyang University with honors in 1998, an M.A. in philosophy from Seoul National University in 2001, and a Ph.D. in philosophy from Brown University in 2014. She worked as a Teaching Assistant for several philosophy courses at Brown University. Her current research interest is in philosophy of action, philosophy of mind, free will, and moral psychology.
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Jaegwon Kim. I first came to be interested in philosophy after reading his works, and his philosophical thoughts and ways of doing philosophy have greatly influenced me ever since. I am especially thankful for his insightful comments, advice, and encouragement during the course of writing my dissertation. I also thank the other members of my dissertation committee, Bernard Reginster and Mary Louise Gill, for their helpful comments and encouragement. And thanks to the philosophy department at Brown University for its generous financial support and for making such a great academic environment.

Personally, I am grateful to my parents and friends for their constant support and prayer. Finally, I would like to give special thanks to my husband, Hongwoo Kwon, for all his love and support.
TABLE OF CONTENTS

CHAPTER 1 Taking Agent Causation Seriously ................................................. 1
CHAPTER 2 Agent Causation and Conceptual Reduction ................................. 23
CHAPTER 3 Agent Causation and Ontological Reduction ................................. 45
CHAPTER 4 The Internal Perspective, Action Explanation, and Agent Causation .... 71
CHAPTER 5 Conclusion: Reduction of Agent Causation ................................. 97
BIBLIOGRAPHY ................................................................................................. 109
1. Introduction

The central problem of philosophy of action is to account for the nature of actions. Actions are fundamentally different from “mere occurrences.” Many events “merely occur” in accordance with laws of nature. They are caused by other events, if caused at all. But human actions seem different. Genuine actions are those that are under agent’s own control. Agents seem to be directly involved in genuine actions. Actions are determined by agents, and so are attributable to them. In short, they are caused by agents.

As actions seem to be fundamentally different from occurrences, there seem to be two fundamentally different kinds of causal relations: event causation and agent causation. The two kinds of causation are distinguished by their relata. Event causation causally relates two (sets of) events, while agent causation relates an agent as a cause on the one hand, and an event as an effect on the other. In order to understand actions and agency, we need to understand what it is for agents to cause certain events. I think that the notion of agent causation has to be central to understanding actions and agency.

The notion of agent causation has been unpopular, to the point that one of its proponents, Randolph Clarke, once admitted, “the notion of agent causation has been
largely abandoned.”¹ I believe that it is because this notion has been associated with a certain philosophical theory about it, which many contemporary philosophers find unpalatable. I have in mind the view held by Roderick Chisholm and others who follow him.² Chisholm begins his classic paper “Human Freedom and the Self” with a summary of a familiar dilemma of determinism and indeterminism, and writes:

“Human beings are responsible agents; but this fact appears to conflict with a deterministic view of human action (the view that every event that is involved in an act is caused by some other event); and it also appears to conflict with an indeterministic view of human action (the view that the act, or some event that is essential to the act, is not caused at all.) To solve the problem, I believe, we must make somewhat far-reaching assumptions about the self or the agent.”³

Chisholm is convinced that moral responsibility is compatible neither with determinism nor with indeterminism. For him, the hypothesis that actions are instances of agent causation is introduced as a “far-reaching assumption,” which is needed to resolve the dilemma.

I believe that the association of the notion of agent causation with this particular theory generated two unfortunate tendencies. First, the notion of agent causation has tended to be regarded as a highly theoretical and metaphysical notion. Agent causation is something that is posited to solve a recalcitrant philosophical problem—that is, the determinism/indeterminism dilemma. Second, the notion of agent causation came to be regarded as a notion that has irreducibility built-in. In other words, agent causation is usually thought to be, by definition, “a sui generis form of causation by an agent that is

² The view I have in mind here is often called “the agent-causal theory”. This view is also held by Taylor (1966), O’Connor (1995, 2000), Campbell (1967), and Clarke (1993, 1996).
irreducible (ontologically as well as conceptually) to event-causal processes within the agent”⁴

But I think that we had better free the notion of agent causation from these associations. First, I believe it is a very intuitive notion based on our everyday practice and experiences, possibly even analytically tied to the concept of action itself. Second, it is one thing that genuine actions are instances of agent causation, and it is quite another that agent causation is a *sui generis* kind of causation that is irreducible to event causation. If one uses the notion of agent causation so as to entail irreducibility, it will not be even possible to state a different philosophical opinion about it coherently. This doesn’t seem to be a fair way to deploy a philosophical notion. In any case, I will use the notion of agent causation in a neutral way with respect to its reducibility.

One reason why I make the last point clear at the outset is that the main thesis of this dissertation is something that would sound contradictory to those who think that agent causation by definition is a kind of causation irreducible to event causation. I believe that we need to be *reductionists* about agent causation, in order to give an adequate account of it. I want to explore whether and how agent causation can be reduced to event causation. This dissertation aims at throwing new light on the problem of agency, by pursuing a reductionist project about agent causation.

In this introductory chapter, I aim to provide a rationale for this project. I will first consider what reasons there are to take agent causation seriously. And then I will examine what reason can be provided against an irreducible kind of agent causation. The main reason, I will argue, lies in the fact that it conflicts with what may be called “the

⁴ Kane (2005), p. 23.
naturalistic or scientific worldview.” I will try to make explicit how the conflict occurs, and claim that some of the influential views about agent causation can be seen as responses to this conflict. I will argue that reducing agent causation to event causation, if successful, is a best way to resolve this conflict.

2. Why Believe in Agent Causation

Why should we believe in agent causation? I think that there are two major reasons. One concerns moral responsibility, and the other phenomenology of agency. I will discuss the first only very briefly, focusing on the second, because the latter has been relatively less discussed in the literature.

2.1. Moral Responsibility

Traditionally, the notion of agent causation has been thought to be intimately tied to that of moral responsibility. As Clarke says, “[a]gent causation … is a condition of the possibility of morally responsible agency.”\(^5\) John Bishop also says, “agents are properly held ethically responsible for outcomes only when those outcomes occur through the exercise of their own control.”\(^6\) That is,

(A) One can be held responsible for his action only if it is cause by himself.

\(^6\) Bishop (2012), p. 54.
This sounds almost like a truism. How can I possibly be held responsible for something that I didn’t bring about? If one finds this doubtful, I suspect that it is because he confuses (A) with the following:

(B) One can be responsible for his action only if it is *not* caused by events.

(A) and (B) are *not* equivalent (unless we assume that agent causation, by definition, is irreducible to event causation). (B) should be controversial, and indeed is denied by many philosophers (who are *compatibilists* about determinism and moral responsibility). However, I don’t think that one can reasonably deny (A).

And as we are indeed responsible for many of our actions, it follows that we cause our actions. I think that this constitutes a powerful argument for agent causation. If one wants to deny that there is agent causation, one will have to deny that we are ever genuinely responsible for our actions, which of course is better to be avoided.

### 2.2. Phenomenology of Agency

The idea that the way in which we experience our actions provides reason to believe in agent causation is quite widespread. For example, C.A. Campbell says:

“Why do human beings so obstinately persist in believing that there is an indissoluble core of purely self originated activity…? There can be little doubt, I think, of the answer. They do so, at bottom, because they feel certain of the existence of such activity from immediate practical experience of themselves.”

---

7 Campbell (1967), p. 41.
Campbell here is saying that we “feel” the existence of agent causation, and that we persistently believe in agent causation on the basis of it. Timothy O’Connor, a leading contemporary proponent of agent causation, makes a similar point.

“[Agent causation] is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favour doing so; it seems to be the case, rather, that I produce my own decisions in view of those reasons.”

Here O’Connor too appeals to our experience of our own activity in order to support agent causation. More recently, Martine Nida-Rümelin says:

“[the] experiences of ourselves as being active and these perceptions of others as active in their doings are deeply incorporated in the way we conceive of ourselves and of others.”

I want to elaborate the reasoning that these philosophers might have in mind, and argue that our phenomenology of agency indeed gives reason to believe in agent causation. In particular, I will try to answer two questions. First, exactly what are the experiences of our agency, and do we really have such experiences? Second, how does it support the belief that there really exists agent causation?

I think that the following passage by Terence Horgan adequately describes the kind of experiences we have when we act:

“How, then, should one characterize the actional phenomenal dimension of the act of raising one’s hand and clenching one’s fingers … It is what-it’s-like-of self as source of the motion. You experience your arm, hand, and fingers as being moved by you yourself—rather than experiencing their motion either as fortuitously moving just as you want them to move,

---

or passively experiencing them as being caused by your own mental states. You experience the bodily motion as generated by yourself.”

That is, there is an experience that you undergo when you act voluntarily, say, raising your hand, and that the experience represents your bodily motion as being caused by yourself. I will state the main claim as follows:

**Phenomenology of agency:** When you act voluntarily, you do not experience the bodily movement as being caused by events (internal or external), but experience them as being caused by yourself.

I find this introspectively compelling. Suppose that your hand is raised. I am sure that if the situation is normal, you can easily tell which of the followings is the case: your hand moves fortuitously, your hand is moved by your strong desires for something, someone else’s pulling up your hand, or you yourself cause it to rise. But how can you tell that? An obvious answer is that you have a certain feeling indicating that it is caused by yourself. You feel as if you caused the movement of your hand. In other words, you experience it as being generated by yourself. As Horgan says, “once you pay proper introspective attention to the features we will describe, you should find such phenomenology ubiquitously present in your own experience.”

One might object as follows: Surely there are visual and kinetic sensations associated with my raising my hand. However, those sensations themselves don’t represent that I caused the movement. Rather, that I caused the movement is what I come to judge or believe on the basis of those sensations and other background knowledge. In

---

11 Horgan, T., Tienson, J., & Graham, G. (2003), p. 323. Tim Bayne also says, “the experience of oneself as an agent is, for many of us, a robust and readily identifiable phenomenon.” (Bayne 2008, p. 185)
other words, you’re mistaking a belief for an experience. But I think that this is very implausible. Such experiences of agency seem to be *independent of what you believe.* Suppose that you are an “eliminativist” about agency, and so that you firmly believe that there is nothing you can cause, all your bodily motions being produced by activities of neurons. As far as I can see, there is no reason why you can coherently believe this sort of thing. But regardless of your belief, when you raise your hand, it will still *seem* to you that the movement is generated by yourself.

Let me try to make our thesis of phenomenology of agency a little more plausible by clarifying what it implies and what it doesn’t.

First, I intend this thesis to be about the experience of *first-personal* agency. That is, it is about the experience that you have about *your own* actions. Some philosophers extend this claim to the third person. For example, Nida-Rümelin claims that we experience other people’s actions in the same way. But this should be controversial, and our thesis of phenomenology of agency does not make any commitment to this.

Second, the thesis concerns only the “representational” aspect of the experience. In philosophy of mind, it is customary to distinguish between two aspects of experience: *phenomenal* and *intentional.* For example, visual experiences are accompanied by vivid phenomenal character, and they also represent the world as being in certain ways. Horgan does talk as if there is a distinctive phenomenal character accompanying experiences of agency, when he says that there is “something it is like.”12 And Carl Ginet also says that there is “actish phenomenal quality.”13 However, one might well doubt whether our

experiences are accompanied by anything like that—apart from visual and kinetic sensations. But this is not a part of our thesis of phenomenology of agency. What is important is that there is some experience that has representational contents.

Lastly, the thesis doesn’t imply that we experience our actions as being not caused by events. In other words, it is not a part of the thesis of phenomenology of agency that we experience agent causation as a sui generis kind of causation. Horgan helpfully puts the point as follow:

“Agentive phenomenology and the phenomenology of state-causation are mutually exclusionary: it is virtually impossible to simultaneously experience a single item of one’s own behavior both as actional and as state-caused. It is easy to make the mistake of inferring, on the basis of the fact that one cannot experience one’s own behavior both as action and as state-caused motion, that no item of behavior can really be both a genuine action and a state-caused bodily motion. … But, psychologically tempting though that inference might be, it is a non sequitur.”

Compare the following two claims:

(1) We do not experience our actions as being caused by events.

(2) We experience our actions as not being caused by events.

The thesis of phenomenology of agency is committed to (1). But (1) does not imply (2), for the simple reason that in general, not experiencing p does not imply experiencing not-P. I think that this point is very important. For the confusion of (1) with (2) might tempt one to think that our phenomenology of actions support that agent causation is a sui generis kind of causation irreducible to event causation. But I think that our phenomenology is neutral on whether agent causation is reducible or not. This is not a question that our phenomenology alone can answer; rather, it is a metaphysical question

concerning the nature of agent causation. In any case, our thesis of phenomenology of agency does not entail that we experience agent causation as a sui generis kind of causation.

I think that with these qualifications, the thesis of phenomenology of agency is very plausible. Now let us turn to our second question. How does the thesis of phenomenology of agency support agent causation? We shouldn’t expect that the thesis entails that there is agent causation, for obviously, there is no logical relation between our having experience as of something and its being real. Nonetheless, I want to claim that the phenomenology does give some support to it. I want to offer two reasons.

First, I want to appeal to an epistemological thesis that seems to be widely accepted among epistemologists. In general, when we have experience as of something, it gives a prima facie reason to believe it.\textsuperscript{15} For example, upon having the experience as of a red circle on the wall, I believe that there is a red circle on the wall. Of course, merely having the experience does not prove that there is really a red circle on the wall, but it at least provides a prima facie reason to believe it. Without such a principle, it is not clear at all how we can avoid wholesale skepticism. Our belief in agent causation based on the experiences of it shouldn’t be exceptions in this regard. It can be defeated by other considerations, but it does give a prima facie reason to believe in agent causation.

Secondly, we can appeal to another general principle concerning the interpretation of representational systems. Nida-Rümelin says:

\begin{flushright}
\textsuperscript{15} E.g., Pryor (2000).
\end{flushright}
“obviously we cannot simply trust the way things appear to us. However, we should be reluctant to accept a theory that implies massive, fundamental and permanent error in the way we experience ourselves ...”

Our experience of agent causation is ubiquitous. If there is no agent causation, it implies that there is massive error “in the way we experience ourselves.” And in general, there is a presumption that other things being equal, we had better avoid attributing massive errors to a representational system.

I should emphasize that the point is not that these considerations prove that there is agent causation. Rather, it is that they shift the burden of proof to those who claim that our phenomenology of agency is always illusory. I think that they will have to do at least two things to defend the error theory. First, they will have to say why we shouldn’t take our phenomenology of agency at face value. Second, they will have to explain why such massive errors in our phenomenology occur.

### 3. Discontents with Agent Causation

In this section, I want to consider what reasons there are to doubt the idea of agent causation. I will consider two reasons. One is that the notion of agent causation is unintelligible, and the other is that it conflicts with “the scientific or naturalistic view of the world.” I will claim that the second is much more worrisome. However, I will claim that it rather gives a reason to take reductionism about agent causation seriously.

---

17 There are philosophers who attempt to answer both questions. E.g., Wegner (2001) and Velleman (1989).
3.1. The Intelligibility of Agent Causation

One reason why some philosophers refuse to accept agent causation is that they find the notion of agent causation unintelligible or at best mysterious. Let me consider the version of such an argument presented by Alvin Goldman:

“It is difficult for me to see how, if [authors like Taylor, Chisholm, and C. A. Campbell] were right, we could ever tell that there are such occasions [of agent causation]. … How do I distinguish my grimacing as an act from my grimacing as a reflex? Or to pose a slightly different problem, how do I tell that I am causing my grimacing and that my grimacing is not totally uncaused? On Taylor’s view, acts are not uncaused; they are caused by agents. But how are we to distinguish absence of causation from causation by agents? Normally we identify causes with the help of regularities, but regularities obtain with respect to events, states, processes, and the like, which are precluded ex hypothesi by Taylor’s view. Thus, the notion of agent-causation unconnected with event-causation is bound to be a mysterious and obscure notion.”

Goldman here presents his argument in epistemological terms. Suppose your hand is raised. How does that happen? Consider two scenarios. First, you cause a bodily movement, and second, the bodily movement just happens without being caused by anything. Goldman claims that we can never distinguish between the two scenarios, and from this he concludes that the notion of agent causation is “mysterious and obscure.”

There is also a metaphysical version of this argument. As Goldman says, “Normally we identify causes with the help of regularities.” But why? Isn’t that because that’s what causal relation consists in? If there is agent causation, then there should be a difference between an action’s being caused by an agent and its just happening. But what

does the difference consist in? The critic will claim that those who believe in agent causation are making a distinction without difference.

I have two responses to this sort of criticism. First of all, these arguments, even if successful, would apply only to the view that takes agent causation to be a *sui generis* kind of causation. So if anything, they show that agent causation needs to be reduced to event causation, not that the notion of agent causation is “mysterious or obscure.” (To be fair, Goldman was aware of this, when he restricted his conclusion to “the notion of agent-causation *unconnected* with event-causation.”)

Second, I don’t think that these arguments are convincing even when applied to agent causation conceived as an irreducible kind of causation. The epistemological version crucially relies on the assumption that we can never tell when an action is caused by an agent and when it just happens. But if what we claimed in the last section is right, this is simply not true. One *can* tell whether his action is caused by himself or not, simply on the basis of his experiences.

I think that the metaphysical version of the argument also relies on a debatable assumption. Chisholm once responded to this concern by saying: “the difference between the man’s causing A, on the one hand, and the event A just happening, on the other, lies in the fact that in the first case but not the second, the event A *was* caused and was caused by the man.”¹⁹ The idea seems to be that we can regard agent causation as, to use Clarke’s phrase, “basic constituents of the universe.”²⁰ What is so bad about it? I suspect that the critic’s worry is based on an epistemological worry. We can know about causal

---

relations only indirectly, by observing regularities. But again, if it is true that we directly experience our actions as being caused by ourselves, this objection will be significantly weakened.

3.2. Conflict with the Naturalistic World View

I believe that a better reason to be suspicious of agent causation comes from what may be called “the naturalistic or scientific view of the world.” For example, David Velleman says:

“our scientific view of the world regards all events and states of affairs as caused, and hence explained, by other events and states, or by nothing at all. And this view would seem to leave no room for agents in the explanatory order.”

John Bishop also says:

“agent-causation has no place—not, anyway, in a contemporary natural ontology. The scientific image of the world is of a vast array of events and states of affairs related according to laws of nature, whether deterministic or statistical. There is no irreducible substance-causation.”

According to these philosophers, it is a commitment to a scientific view that all natural events are caused by, and be explained in terms of, other events and states that excludes (irreducible) agent causation. Let us consider more closely how this reasoning works.

Sciences aspire to causally explain all phenomena or events by referring to scientific laws, either deterministic or statistical. And as scientific laws connect different types of events, we should assume that every event, if it can be explained at all, can be

completely explained by referring to other events and laws subsuming them. This view naturally suggests the following:

**The Naturalistic View of the World**: Every event E, if it has a cause at all, has another event C (or a group of events) as its complete cause.

I need to make clear what I mean by saying that event C is a “complete cause” of another event E. It does not mean that C is a *sufficient* cause, in the sense that C causally *necessitates* E. As observed by many philosophers, event C may qualify as a cause of E even if the occurrence of C does not necessitate, but only increases the probability of the occurrence of E. In other words, the scientific view of the world should be compatible with “probabilistic causation.” But even when C only increases the probability, it can be a complete cause of E in the following sense: C provides a fully satisfactory explanation of why E occurred, and no further explanation is needed or possible.

Granting that the naturalistic worldview as formulated above is compelling, how does it rule out agent causation? Both Velleman and Bishop, in the above passages, seem to take it for granted that the naturalistic view directly rules out agent causation. But as Maria Alvarez and John Hyman point out in the following passage, things are not so straightforward.

“if it is true that our scientific view of the world regards all events and states of affairs as caused, and hence explained, by other events and states, or by nothing at all, it does not follow that there is ‘no room for agents in the explanatory order.’ … if all events are caused by other events and states, or by nothing at all, some events may also be caused by agents. It is
the view that only events can cause events which seems to leave no room for agency …”23

Alvarez and Hyman distinguish two possible formulations of “our scientific view of the world.” One is that “every event that has a cause is caused by another event, or by nothing at all”24 (which is virtually equivalent to our formulation), and the other that “only events can cause events.” And the latter is clearly not implied by the former. According to them, it is only when we assume the second version that agent causation is ruled out.

However, if we combine the naturalistic view with a very plausible metaphysical principle, it does seem to entail that actions are not caused by agents. What I have in mind is what Jaegwon Kim calls the causal exclusion principle:

**The Exclusion Principle**: For any event that has a cause, there cannot be two complete and distinct causes.25

There may be cases where two distinct causes “overdetermine” a single effect—as when a fatal heart attack and a fatal stroke cause a person’s death exactly at the same time. But surely there is something odd with the idea that this sort of overdetermination can systematically occur.

Kim introduces this principle in relation to the mind-body problem. He asks, “*Given that every physical event that has a cause has a physical cause, how is a mental cause also possible?*”26 Similarly, we may ask: Given that every action that has a cause

__________________________

has an event-cause, how is agent causation also possible? We can present the argument more precisely, as follows:

(1) By the naturalistic view, every action, if it has a cause, has another event as its complete cause.

(2) By the exclusion principle, if an action has a complete cause X, then there is no other complete cause Y that is distinct from X.

(3) Therefore, no action is caused by an agent.

This argument seems to be valid. I find the exclusion principle, when it is correctly understood, extremely plausible. So assuming that the naturalistic view of the world is true, agent causation, if it is really distinct from event causation, is on the verge of being excluded.

4. Responses

Our situation is this. We have good prima facie reasons to believe in agent causation. But a commitment to agent causation apparently conflicts with the naturalistic view. How can one respond to this conflict?

One option is to accept the conclusion of the above argument, and give up agent causation. For example, the psychologist Daniel Wegner, in his recent book *The Illusion of Conscious Will* can be interpreted as taking such a position. He suggests, on the basis of empirical research, that we may uncover the psychological or physiological mechanisms that cause our actions that we think we ourselves cause (or in his terminology, our “conscious will” causes). Wegner says:
“The idea of conscious will and the idea of psychological mechanism have an oil and water relationship, having never been properly reconciled. One way to put them together … is to say that the mechanistic approach is the explanation preferred for scientific purposes but that the person’s experience of conscious will is utterly convincing and important to the person and must be understood scientifically as well. The mechanisms underlying the experience of will are themselves a fundamental topic of scientific study. … This means, though, that conscious will is an illusion.”27

Here Wegner seems to be implicitly appealing to the exclusion principle. As explanations of actions in terms of “conscious will” and those in terms of psychological mechanism are incompatible with each other, he opts for “the mechanistic approach,” and declares that “conscious will is an illusion.”28

On the other extreme, there is Chisholm. As we saw in the beginning of this chapter, he takes for granted that agent causation is a 

\textit{sui generis} kind of causation. And he clearly observes that this is incompatible with the naturalistic view, and opts to deny the latter to save agent causation. He says:

“in one very strict sense of the terms, there can be no science of man. If we think of science as a matter of finding out what laws happen to hold, and if the statement of a law tells us what kinds of events are caused by what other kinds of events, then there will be human actions which we cannot explain by subsuming them under any laws.”29

However, many will find this just too extreme.

There is a more moderate option that tries to save agent causation by weakening the naturalistic view, instead of disowning it altogether. As I understand it, Randolph Clarke’s position can be understood as such a position. He says:

\begin{flushright}
\textbf{27} Wegner (2002), p. 2. \\
\textbf{28} “Hard determinism” about free actions can be also regarded as a similar position. \\
\textbf{29} Chisholm (1964), p. 35.
\end{flushright}
“since an agent-causal thesis does not require that there be any gaps in chains of event causes, agent causation does not undermine the predictive and explanatory significance of event causes. Indeed, agent causation is consistent with its being the case that probabilistic laws of nature apply as thoroughly to human beings and their behavior as such laws apply to anything else.”

Unlike Chisholm, Clarke doesn’t deny that we can explain human actions by subsuming them under laws. But he does deny that science can provide complete explanations of all genuine actions. At best, it can only partially explain actions in terms of probabilistic laws. To be complete, according to Clarke, an explanation of a genuine action has to refer to the agent himself as another partial cause.

So all these positions attempt to save one intuition at the cost of sacrificing the other. But I believe that there is another option that enables us to retain both of them, to which we now turn.

5. The Reductionist Project

I want to claim that the conflict is best resolved by taking a reductionist position. Let me first try to motivate this position by drawing attention to a parallel between the mind-body problem and the problem at hand, which I briefly mentioned earlier in passing.

We find it intuitively obvious that mental states and events exert causal influence on the physical events. Many of us also subscribe to the doctrine of “physicalism,” according to which if a physical event has a cause, it has a physical cause. However, to those who are committed to both of these, an exclusion problem occurs: “Given that

---

every physical event that has a cause has a physical cause, how is a mental cause also possible?” The mental as distinct from the physical is on the verge of being excluded.

In response to this problem, some may want to be eliminativists about the mental realm (which corresponds to Wegner’s position). Others may prefer to reject physicalism (which corresponds to Chisholm’s position). But many philosophers would opt for some forms of reductionism. Why is this a solution? Because the exclusion principle only rules out two distinct and complete causes. If the mental is reducible to the physical, the exclusion principle may not apply.

Then why not do the same for agent causation? In other words, we may be able to resolve the conflict by reducing agent causation to event causation. We say that a certain action is caused by the agent, and also say that the same action is caused by some psychological or physiological mechanisms. But perhaps they are merely different ways of describing the same phenomena. An action’s being caused by an agent and its being caused by certain events may not be really distinct, and so the exclusion principle may not apply.

The merit of this position, assuming that it can be successfully completed, is obvious. All other positions we considered in the last section sacrifice some of our strongly held beliefs, and the cost of doing so is not negligible. If we can save those beliefs, it will be a great advantage of the reductionist position over the alternatives, by a simple cost-benefit analysis. Moreover, on the face of it, there is no obvious reason to think that such reduction is not feasible.
Earlier in this chapter, I voiced complaints about the way in which Chisholm framed the problem of agent causation. However, it is interesting that he, in his later writing poses the problem in a quite different way:

“Do you agree that such statements tell us something about an agent and one or more events that that agent caused to happen? I would hope that they might agree to this much: ‘I raise my arm’ tells us, with regard to my arm going up, that I made it happen; ‘Jones killed his uncle’ tells us, with respect to the death of Jones’s uncle, that Jones brought it about; and ‘Smith painted the back fence’ tells us that Smith caused it to happen that parts of the back fence were covered with fresh paint. … I would say that if Jones brought it about, then an agent brought it about, and if an agent brought it about, then we have a case of ‘agent causation.’ … So the issues about ‘agent causation,’ as I say, have been misplaced. The philosophical question is not—or at least it shouldn’t be—the question whether or not there is ‘agent causation.’ The philosophical question should be, rather, the question whether ‘agent causation’ is reducible to ‘event causation.’”  

I am inclined to agree with what Chisholm says here. Agent causation is something intuitively obvious and undeniable, and we need to take agent causation seriously. The question is not whether there is agent causation or not, but whether it can be reduced to event causation.

6. Plan

The merit of the reductionist position, of course, can be had only if the reductionist project can be successfully completed. The chapters to follow are devoted to exploring whether and how this project can be completed. Here is my plan.

In Chapter II, I consider whether agent causation can be *conceptually reduced* to event causation. My answer is negative. I examine various reasons adduced for conceptual irreducibility. I will suggest that the real reason is that agent causation and event causation belong to different “conceptual schemes” or “perspectives,” postponing discussion of what these amount to until Chapter IV.

In Chapter III, I turn to the idea that agent causation is not conceptually, but *ontologically reducible* to event causation. For this, I try to make clear what ontological reduction should be. I claim that reduction of agent causation to event causation should conform to the so-called “functional model of reduction.” I favorably consider Velleman’s proposal in this context. I claim that functional reduction of agent causation is promising, but also identify some problems that need to be addressed.

Chapter IV discusses an idea that emerges from the discussion in Chapters II and III. It is that agent causation is something visible only from agents’ internal perspective, while event causation can be seen from the external point of view. I seek to explain what these different perspectives amount to. I claim that they lie in two different practices of *explaining* actions, and that the concept of agent causation is tightly connected to one particular kind of explanation of actions, which I call “normative explanation.”

The discussion in Chapter IV will provide materials to solve the problems that Chapter III leaves unsolved. In Chapter V, I claim that the role of agents in producing actions needs to be found in the context of normative explanations. I also make a rough suggestion about what agent causation is reduced to.
CHAPTER 2

Agent Causation and Conceptual Reduction

1. Introduction

In the last chapter, I argued that there is good reason to be a reductionist about agent causation. In this chapter, I begin to consider whether reduction of agent causation to event causation is possible.

What does it take to reduce agent causation to event causation? As we saw in the last chapter, Roderick Chisholm took the central philosophical question about agent causation to be whether it is reducible to event causation. And for Chisholm, the philosophical question about reduction comes down to this:

“[I]f we have good reason for believing that Jones did kill his uncle, then the philosophical question about Jones as cause would be: Can we express the statement ‘Jones killed his uncle’ without loss of meaning into a set of statements in which only events are said to be causes and in which Jones is not said to be the source of any activity? And can we do this without being left with any residue of agent causation—that is, without being left with some such statement as ‘Jones raised his arm’ wherein Jones once again plays the role of cause or partial cause of a certain event?”  

Chisholm thinks that the question of reduction is the question of conceptual analysis or translation. It is a matter of expressing statements about agent causation into statements of event causation “without loss of meaning.” More recently, E. J. Lowe too identifies

---

the reductionist thesis with the claim that “any statement of agent causation can be analyzed in terms of a statement in which no notion of causation other than that of event causation is employed.” And it is because they think that this sort of reductive analysis is impossible that they conclude that we need a *sui generis* kind of agent causation.

However, on the face of it, there is something odd with the way these philosophers understand reduction. It is true that reductive analysis in this sense would get us reduction of agent causation to event causation. However, it seems to me wrong to identify reduction with conceptual analysis. For conceptual reduction in this sense is by no means the only model of reduction, let alone the most popular one among contemporary philosophers. For example, if we want to reduce pain to a certain physiological state, or heat to molecular kinetic energy, what we want is not a conceptual analysis of the former in terms of the latter, but something different. Is there any reason to think that agent causation is special in this regard?

I will set this issue aside in the present chapter. I think that it is an important and substantial claim that conceptual reduction of agent causation is impossible, especially because there have been attempts for such reduction. So I will take up the issue of conceptual reducibility of agent causation in this chapter. I will argue that such reduction is indeed impossible, and also try to identify the reason behind irreducibility.

I will start with simple proposals to reductively analyze agent causation in terms of event causation. And then I will consider the so-called “causal theory of action” and its variants construed as proposals for conceptual reduction of agent causation. After that, I

---

will try to identify a more fundamental reason that we shouldn’t expect agent causation to be conceptually reducible to event causation.

2. Initial Attempts

We say that in doing something, an agent causes something to happen, makes it happen, or brings it about. And we also say things like “A tree caused the roof of my car to collapse.” That is, we say that certain inanimate things cause certain effects. But in the case of non-agent subjects, such a statement almost always turns out to be an elliptical way of saying that some event involving the thing causes the effect. For example, everyone will agree that when we say that a tree caused the roof of my car to collapse, it is a mere elliptical way of saying something to the effect (say) that the tree’s falling caused the roof’s collapsing.

This suggests a simple analysis of agent causation in terms of event causation. A tree can be said to cause the collapse, in that some event involving the tree causes it. Then simply by generalizing this to agent causation, we get the following:

(A) Agent A causes event E if and only if some event C (or a class of events) involving A causes E.

But it is easy to see that this fails as an adequate analysis of agent causation. For example, a person’s weighing 120lb causes reading “120lb” on a scale. This satisfies the right side of (A); that is, an event involving that person (the person’s weighing 120lb) causes the reading on the scale. But of course, this shouldn’t count as a case of genuine action.
(There may be a sense in which we can truly say that the person caused the reading on the scale. But obviously this doesn’t capture the sense of agent causation we are interested in.)

What this shows is that for a successful analysis of agent causation, we need to restrict \( C \) in the right side of the analysis to a special class of events involving agents. What could that be? In this context, let’s consider the following suggestion from Donald Davidson:

“If I poison someone’s morning grapefruit with the intention of killing him, and I succeed, then I caused his death by putting poison in his food, and that is why I am the agent in his murder. When I manage to hurt someone’s feelings by denigrating his necktie, I cause the hurt, but it is another event, my saying something mean, that is the cause of the hurt. … The notion of cause appealed to here is ordinary event causality, the relation, whatever it is, that holds between two events when one is cause of the other. For although we say that the agent caused the death of the victim, that is, that he killed him, this is an elliptical way of saying that some act of the agent—something he did, such as put poison in the grapefruit—caused the death of the victim.”\(^{34}\)

What Davidson is considering is the following suggestion:

(B) Agent A causes event E if and only if some act of A causes E.

This seems right in many instances of genuine actions.\(^{35}\) We may say things like “A person caused the nuclear war.” But this is certainly analyzable into the statement that a certain behavior of the person, such as pushing the button of a bomb control, which is an event, causes the nuclear war to start.

\(^{34}\)Davidson (1971), pp. 48-9.

\(^{35}\)More precisely, (A) seems to be true of all and only non-basic actions. Non-basic actions are actions that an agent performs by performing some other actions. On the other hand, basic actions are actions that an agent performs without performing something else.
However, it is not difficult to see that this fails as a *reductive* analysis. We said above that event C in (A) needs to be restricted appropriately. The above proposal in effect is restricting the class of events to *acts*, whose examples are my putting poison, my denigrating someone’s necktie, and my saying something, etc. But they are simply instances of genuine actions that are caused by agents. In order for the conceptual analysis to be successful, we need to specify “acts” without referring to the notion of agent causation. But that’s precisely what we have to do to conceptually reduce agent causation in the first place. (B) may help to isolate what an agent really causes from causal relations between events. But when it comes to reductive analysis of agent causation, we haven’t made any progress.

From our discussion of the two simple suggestions, it becomes clear what needs to be done to conceptually reduce agent causation to event causation. We need to identify a special class of events that play the role of agents, and this should be done in a non-question-begging way, that is, without referring to the notion of agent causation. In the next two sections, we will consider more sophisticated proposals, which attempt to restrict C in (A) to a class of *mental events or states*. But I will argue that all these positions face similar problems as the ones that inflict (A) and (B). That is, each position either fails to yield an extensionally equivalent analysis, or fails to be genuinely reductive.

3. The Causal Theory of Action

What is called “the causal theory of action” has been a predominant view in the philosophy of action and in the philosophy of mind. According to this view, actions are bodily movements that are caused in the right way by some appropriate mental events
and states of an agent such as desires, beliefs, or intentions, which rationalize and justify his behavior. For example, suppose that you raise your arm. That movement may be explained as follows: You want to give a signal and you believe that you can give a signal by raising your arm. And that’s why you raise your arm. If these two mental attitudes jointly cause the bodily movement of my arm’s rising in an appropriate way, according to the causal theory, this is an instance of genuine action.\(^\text{36}\)

The causal theory of action purports to give a theory of action and agency. But as we have been assuming that the core of a theory of action is to give an account of agent causation, we may construe the causal theory as an attempt to conceptually reduce agent causation to event causation. The idea is that an agent’s causing something consists in its being caused by the agent’s belief and desire.

I think that there are at least two reasons why this line of approach may look promising. In the last section, I said that we needed to identify a class of events that play the role of agents. Events that could play this role (unlike the event of a person’s weighing 120lb) should be those that are central to agents’ identity. David Velleman writes:

“[T]he events that it picks out in the causal history of my behavior are closely associated with my identity, and the causal operations of these events consequently implicate me, at least to some extent. What I want and what I believe are central features of my psychology, which is central to my nature as a person. My wanting and believing are therefore central features of me, and whatever they cause can be regarded as caused by me, in some sense.”\(^\text{37}\)

\(^{36}\) Davidson is a leading proponent of this view. See Davidson (1980). This view is sometimes called “the standard story of action”. See Velleman (2000).

If an agent’s causing something is conceptually equivalent to some events’ causing it, the events must be those that are essential, or at least central, to the agent’s identity. A lot of different things must be involved in being an agent, but surely, some psychological features will play predominant roles. In particular, as Velleman says, what an agent believes and desires seem to be “central features” of the agent. If this is right, then it seems to be a very natural move to analyze an agent’s doing something in terms of beliefs and desires causing the behavior.

The second reason comes from the idea that an agent’s causing something is closely related to the idea of the agent’s acting for reasons. Intuitively, it seems that an action is attributable to an agent when the agent performs the action for reasons, or in responding to reasons. Think of a typical case of genuine action where an agent makes something happen. The agent will consider various reasons for or against alternative actions, and on the basis of the calculation, he will determine what to do, and brings about the consequence that he judges to be the most appropriate. In short, as Velleman puts it, “whatever we do for reasons is … of our making.”[^38] But what kind of things are the reasons that are involved in this sort of deliberation, and how should the agent be related to those reasons in order for him to act for those reasons? An influential view has it that those reasons consist in the agent’s desires and beliefs, and that the agent performs an action for those reasons if the agent is caused to move by those mental states.[^39] An analysis of agent causation in terms of an agent’s belief and desire can be seen as a

[^39]: For this view, see Davidson (1963). Reasons in this sense are “explanatory reasons” rather than “justifying reasons.” In Chapter IV, I will claim that what is really central to agent causation is not explanatory reasons in this sense, but justifying reasons.
promising attempt to analyze what it is for the agent to do something for reasons, which
in turn seems to be conceptually related to agent causation.

Now let us formulate the causal theory’s analysis of agent causation in the following way:

(C) Agent A causes event E if and only if E is caused in the right way by A’s belief and desire that rationalize it.\textsuperscript{40}

I will call this “the belief-desire analysis” of agent causation. Some remarks are in order. First, the belief and desire that are supposed to cause E have to be those that rationalize the behavior. In other words, they are the mental states that constitute the agent’s reasons for acting. This should be a matter of some appropriate relations between contents of beliefs and desires, on the one hand, and the behavior on the other. In typical cases, a desire for D, and the “instrumental belief” to the effect that by making movement M, one can achieve D, will rationalize M.

Second, (C) says that E has to be caused “in the right way” by the desire and belief that rationalize it. This is included to avoid the so-called the problem of “deviant causal chain.” There may be a case in which a behavior is caused by an agent’s belief and desire that rationalize it, but doesn’t count as an action, because the causal process from the desire and belief to the behavior proceeds in a deviant or abnormal way. Consider Davidson’s climber example:

\textsuperscript{40} In a more improved version, an intention is added to the picture as playing an intermediating role between desire/belief pair and a bodily movement. In this version, desires and beliefs cause intentions, and intentions in turn cause a bodily movement. But here I take the causal theory as “the desire-belief model” for the sake of simplicity.
“A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope, he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally”.

In this case, the climber’s belief and desire rationalize his loosening the hold. Moreover, these states are jointly supposed to cause his loosening it. But the causal chain leading from the desire and belief to the action runs through a certain abnormal process. The causal theorists obviously don’t want to include this sort of behavior as a case of genuine action. I am not sure what is the best way to spell out what it is to be caused “in the right way”, but let me grant that the causal theorist somehow can specify this condition in a non-question begging way; that is, not by appealing to the notion of agent causation, or something of that sort. (If this can’t be done in a non-question begging way, so much worse for the belief-desire analysis.)

Now even if this can be done, I want to argue that (C) fails as an adequate analysis of agent causation. There are counterexamples to this analysis. One example I want to consider is Harry Frankfurt’s famous example of an “unwilling addict.” Suppose that a drug addict has a strong and irresistible desire to take a drug. We may assume that he is aware of such a desire, feeling a strong impulse for a drug. But the addict may be “alienated” from that desire, in that he wants to resist the desire. Now suppose that this desire, with a certain appropriate belief, drives the addict to take the drug. There is no reason to think that the addict’s behavior is being caused in a deviant way as in the case

41 Davidson (1980), p. 79.
42 Some philosophers seem to think that the problem of deviant causal chain is a central problem that the causal theory’s analysis of agent causation has to solve. For example, see Bishop (1989), Chs. 4 & 5.
of Davidson’s climber example. So the addict’s behavior of taking the drug satisfies the right side of (C). Still, many would think that this behavior doesn’t count as a genuine action that the agent does for a reason, or an action that is caused by him. The addict will feel as if he is driven by an external force that he can’t control. As Frankfurt says, in such a situation, the addict becomes “a helpless bystander to the forces that move him.”43

The case of the unwilling addict involves the pathological condition of addiction. But there are more banal cases that show the similar point. Here is one from Velleman:

“Suppose that I have a long-anticipated meeting with an old friend for the purpose of resolving some minor difference; but that as we talk, his offhand comments provoke me to raise my voice in progressively sharper replies, until we part in anger. Later reflection leads me to realize that accumulated grievances had crystallized in my mind, during the weeks before our meeting, into a resolution to sever our friendship over the matter at hand and that this resolution is what gave the hurtful edge to my remarks.”44

In this case too, the agent’s behavior of raising his voice, we may assume, is caused by his desire to break up friendship, with an appropriate belief. But there seems to be a sense in which the agent is alienated from his desire. It is assumed that the agent wasn’t even aware that he has such a desire.45 He didn’t choose to act on that desire, and the desire takes effect on its own, as if it is an external force. As Velleman puts it in a first person, in this situation it is “my resentment speaking, not I.”46

45 Velleman assumes that the agent wasn’t aware of the desire at the time of the behavior. But I don’t think that this is really essential. We may assume that the agent came to know that he had such a desire from his past dispositions to act. But still the counterexample will go through, if we assume that the agent doesn’t want to be motivated by such a desire in that occasion.
46 The above examples are cases where agents are driven by desires from which the agents are alienated. But I think that a similar thing can happen with respect to beliefs too. For example,
I think that the lesson from these cases is this. In the case where an agent can be said to cause an action, the action is not simply directly caused by his belief and desire. The agent himself chooses to be motivated by those desires and beliefs. But the belief-desire analysis, by classifying any behavior appropriately caused by desires and beliefs, omits this essential element.

4. Amending the Causal Theory

In this section, I want to examine attempts to amend the belief-desire analysis. The key idea is provided by Frankfurt. As we saw, there is a sense in which the addict is alienated from the desire that motivates him to act. According to Frankfurt, the addict’s act of taking the drug is not attributable to him because the agent doesn’t reflectively endorse, or identify with, the desire that moves him to take the drug. So in his view, the desire that an agent doesn’t endorse and the action motivated by the desire are less central to him, while the desires that he reflectively endorses and the action motivated by it are truly his own. That is, what makes an action fully his own is the agent’s reflective endorsement of the desire that motivates him to act.

If this is right, then the condition that we need to add to the belief-desire analysis is that an agent should endorse the desire that motivate the agent to move. So we can amend the belief-desire analysis in the following way:

---

suppose that an agent obsessively believes that washing his hands as often as possible will cure his cold, and that this belief with the desire to be cured of the cold, derive him to wash his hands. We would be reluctant to see this as an action caused by himself.

(D) Agent A causes E if and only if E is caused in the right way by A’s belief and
desire that rationalize it, and A endorses the desire.

The reason why the addict’s behavior of taking the drug is not an instance of genuine
agency is that he is caused to move by a desire that he does not endorse. That is, the
addict’s behavior fails to satisfy the second condition of (D).

Something like this may be on a right track. But it by itself won’t do. What we
want is a *reductive* analysis of agent causation in terms of event causation, and the
analysans should not contain any concept that is only explicable in terms of agent
causation and other notions related to it. But this may be very doubtful as to (D). What is
it for an agent to endorse his desire? Isn’t it just for the effectiveness of the desire to be
under the agent’s control? If something like this is all we can say about what
endorsement is, then we haven’t made much progress in conceptually reducing agent
causation to event causation. We need to be able to specify a mental state or event that
plays the role of endorsement in a non-question-begging way.

Frankfurt was aware of this problem, and makes a specific proposal about what
constitutes an agent’s endorsing his desire. He claims that certain second-order desires
(or what he calls “second-order volitions”) can do the job. It is a desire for a first-order
desire to be effective. The drug addict doesn’t endorse his first-order desire, in that he
does not want his first-order desire for the drug to motivate him. Likewise for the case of
“raising his voice.” Although the agent’s desire for breaking up friendship causes his
raising voice, he doesn’t endorse this first-order desire, in that he is not even aware of

that desire. (But even if he were aware of the desire, he would not still endorse it, because he doesn’t want to be motivated by the desire in that particular occasion.)

However, this analysis faces a challenge. The reason why the original belief-desire analysis fails is that an agent can be alienated from the motive that causes him to act. The agent’s having a relevant second-order desire is supposed to keep the agent from being alienated from it. But, intuitively, in order for the second-order desire to be able to do this, it should be something from which the agent cannot be alienated. But an agent can be also alienated from her second-order desires. After all, a second-order desire is simply a desire on a par with other desires in relevant respects. If it is possible for an agent to have a first-order desire that he does not endorse, why can’t he have a second-order desire that he does not endorse? Suppose that the addict is an obsessive hedonist, and has a second-order desire (which he does not endorse) to be motivated by his first-order desires. In such a case, the addict’s behavior will not count as an instance of agent causation, even though he has the second-order desire. Obviously, adding a third-order desire will not help, as the agent in principle can be alienated from the third-order desire too.

There is a different proposal that purports to improve the basic insight of Frankfurt’s view. Gary Watson, in effect, claims that the desires that an agent endorses are those that reflect the agent’s values.49 The unwilling addict is alienated from his desire, because his crave for the drug does not reflect his value. Perhaps although the addict wants to take the drug, he at the same time takes that to be wrong, and something to be avoided. And it is in that sense that the addict does not endorse his desire for the drug.

49 Watson (1975).
This may sound attractive, and there may be something ultimately right about it. However, this proposal, as a reductive analysis of agent causation, faces a similar challenge. What is it to value something? Suppose that it is a matter of believing something (perhaps to the effect that doing something is good), or some combination of desires and beliefs. If this is right, then this suggestion will face exactly the same challenge that Frankfurt’s view faces. One can be alienated even from values.⁵⁰ On the other hand, suppose that valuing something is a distinctive attitude yet to be specified. Then the suggestion will face the problem that inflicted the original proposal (D). Unless we specify what that attitude is in a non-question-begging way, we haven’t reduced agent causation to event causation.

I considered the belief-desire analysis and some amendments of it, and argued that they are lacking as reductive analyses of agent causation. I should emphasize that my point is not that the causal theory of action and its variants are wrong. On the contrary, I believe that the causal theory successfully identifies an important class of human behaviors. Moreover, I believe that it is a theory that can adequately explain many cases of genuine actions.⁵¹ But it is a different matter whether it can serve as a reductive analysis of agent causation.

Thus far we considered individual proposals of conceptual reduction of agent causation. But there are philosophers who think that there is a more fundamental reason

---

⁵⁰ For a criticism of Watson along this line, see Velleman (1992), p. 134.
⁵¹ I will have much more to say about the relation between the way in which we explain our actions causally, and the way in which we explain our actions by referring to agent causation in Chapter IV.
to think that any such reductive analysis isn’t forthcoming. I turn to that in the next section.

5. The Problem of Disappearing Agency

Many philosophers who take agent causation seriously seem to think that conceptual analysis of agent causation in terms of event causation is simply impossible. The following are some of the typical expressions of the intuition.

“It is futile to attempt to explain conduct through the causal efficacy of desire— all that can explain is further happenings, not actions performed by agents. The agent confronting the causal nexus in which such happenings occur is a helpless victim of all that occurs in and to him. There is no place in this picture … even for conduct.”

“The essential source of the problem is a view of persons and their actions as part of the order of nature, causally determined or not. That conception, if pressed, leads to the feeling that we are not agents at all that we are helpless and not responsible for what we do. … my doing of an act … seems to disappear when we think of the world objectively.”

“Conceiving of action as the result of a chain of events that are causally connected removes the agent from the picture altogether. Being the conduit for a causal chain is a passive affair; being an agent is being active. The former cannot possibly exhaust what is involved in the latter.”

“If we analyze it [action] as an event caused in some particular way by events and states of an agent, there is always room to ask for a justification for equating those particular events or states with the agent herself. There is always possibility, it might seem, that the agent could have been the passive victim of such events, and it follows, or at least it seems to follow, that no causal theory can serve to capture the crucial aspect of agency, the

54 Enç (2003), p. 3.
sense in which agents are *active*—as opposed to passive—in the production of their actions.\textsuperscript{55}

“This proposal [for amending the causal theory] is not addressed to the standard story’s real difficulties. If human actions cannot be located among states and events viewed as part of the flux of events in nature, then introducing another state into that same flux could never be a recipe for bringing them in”.\textsuperscript{56}

The common objection voiced in these passages targets any possible attempt to analyze agent causation to event causation. It is as if they are saying: “Bring whatever analysis you prefer, however complex. Even without knowing details, I know it is doomed to failure. For whatever events caused the action, the agent is bound to be left out from the picture as the passive victim.”

But precisely what is supposed to be the problem? The critics say that in the reductionist picture, the agent merely serves as a locus in which the mental states and events are taking place, or just “a passive victim” of the causal forces of the states and events. The agent as a source of his action is left out of the picture. It seems that the critics are getting at something, but it is not easy to see what the real force of the objection is.

On the face of it, there seems to be something unfair with this objection. What reductionists try to do need not be anything like “eliminative” reduction. That is, their point need not be that some events do the work *instead of* the agent; if they said so, it would certainly leave out the agent. Rather, their point is that being an agent simply consists in being in certain mental states and events of the agent, or at least that those

\begin{flushleft}
\textsuperscript{55} Yaffe (2000), p. 112.  \\
\end{flushleft}
mental states are “central to” the agent’s identity. And as far as those states and events cause the action, there is a sense in which the agent is doing the work. In a way, then, what the above passages are saying seems to be just an outright denial of the main point of the reductionist’s view, without any substantial argument.

The objector may protest that certain mental states and events of an agent being “central to” his identity is not enough to capture the sense of the agent’s being a source of his action. They invoke the thought that an agent or a self is not identical with any arrangement or a sum of mental states and events. For examples, A.I. Melden says, “I am not any one of these [mental antecedent] factors, nor all of them,” implying that these mental states’ causing an action doesn’t add up to my doing it. And Richard Taylor also writes in a similar vain, “it is plain that, whatever I am, I am never identical with any such event … or state as is usually proposed as the real cause of my act… hence, if it is really and unmetaphorically true … that I sometimes cause something to happen, this would seem to entail that it is false that any event … or state not identical with myself should be the real cause of it.”

Of course, these philosophers are right in pointing out that an agent or a self is not strictly identical to (a sum of) mental events—saying that would be a “category mistake,” as they are clearly different kinds of entities (one is “an object” or “a whole being” and the other “events or states”). But it is not clear whether this is a decisive consideration against the viability of conceptual reduction. For when the reductionists maintain that the agent’s causal role is reduced to the causal role of a certain mental states and events, they

\[58\] Taylor (1966), p. 111.
don’t necessarily commit themselves to the “literal” view that an agent or person is reduced to, or identical with, a sum or group of mental states and events. That is, when it is claimed that “I do something” or “I cause something to happen” can be reduced to the statement that my mental states cause it, this doesn’t have to mean that the word “I” can be substituted, without the change of meaning, by a certain expressions referring to a sum of certain mental events and states. Rather, their point may be that the whole statement, “I do something” may be paraphrased without loss of meaning to statements about event causation.

Should we conclude then that the objectors’ concern is groundless? Some philosophers may be tempted to conclude so. For example, Markus Schlosser says that “its proponents have not produced a single argument to support their case, and they have certainly not identified a philosophical problem.”\textsuperscript{59} But I find such a response also unsatisfactory. If our concern is conceptual reduction, then we should not ignore the sense of uneasiness many philosophers feel toward the attempts for conceptual reduction.

6. Subjectivity of the Concept of Agent Causation

In this section, I want to suggest that what is behind the intuition of “disappearance of agency” is that there is something essentially “subjective” about the concept of agent causation. What I will say in this section will be incomplete, and will be discussed at length in Chapter IV.

\textsuperscript{59} Schlosser (2010), Sec. 3.
The point explicitly comes up in the passage we quoted from Nagel in the last section. Let me quote another passage from him:

“Something peculiar happens when we view action from an objective or external standpoint. Some of its most important features seem to vanish under the objective gaze. Actions seem no longer assignable to individual agents as sources, but instead components of the flux of events in the world of which the agent is a part. … That conception, if pressed, leads to the feeling that we are not agents at all, that we are helpless and not responsible for what we do. Against this judgment the inner view of the agent rebels.”

Nagel seems to think that there is something essentially “subjective” about agent causation (although he himself doesn’t use the notion of agent causation in this way). It is only when we occupy the internal perspective that we appear as real authors of our own actions. He does not deny that we can ascribe agent causation to others, but says that it is by occupying the same perspective “sympathetically with regard to the actions of others.” And when we “view ourselves from an objective or external standpoint”, he says, agents as sources disappear.

I want to suggest that the concept of agent causation is an “essentially subjective” concept. It is a concept that is essentially tied to the way in which we experience our own actions (as we saw in the last chapter), and also essentially tied to the way in which we explain and understand our own actions (as we shall see in Chapter IV). But there is nothing subjective about the concept of event causation. The concept of agent causation and that of event causation seem to belong to, as it were, completely disparate vocabulary or “conceptual schemes.” Suppose that some analysis of agent causation of the following sort is proposed.

\[60\] Nagel (1986), p. 110.
(E) Agent A causes event E if and only if … [some statement about certain events’ causing E].

Can this really be true as a matter of conceptual analysis? The problem is that the left side is essentially first-personal, while the right side is not. So any analysis is bound to leave out something that is essential to the concept of agent causation. As Nagel says in a different context, “the reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view.”

It needs to be made more precise in what sense the concept of agent causation is essentially subjective. I will address this issue in Chapter IV. For now, let me try to make this idea a bit more intuitive with help of an analogy.

Take the concept of colors, say, red. There is a debate about the nature of colors. Some philosophers think that it is a certain dispositional property, and others claim that it is a physical property. But I think that there is a virtual agreement that it is impossible to conceptually analyze the concept of red in physical terms. Why so? Because the concept of red is essentially tied to the way we experience it. Red is a property that typically tends to produce a certain experience to us. By trying to analyze it in physical terms, we will necessarily leave out what is essential to the concept of red. So we can’t translate statements about color red to statements about physical properties.

Similarly, the reason why many philosophers strongly feel that agent causation is conceptually irreducible to event causation is that the concept of agent causation and the concept of event causation belong to disparate conceptual schemes. By trying to translate

\[\text{\textsuperscript{61}}\text{Nagel (1974), p. 437.}\]
statements about agent causation into statements about event causation, we leave out an essential feature of the concept of agent causation—that is, its subjectivity. It seems to me that something like this successfully brings out the source of discontents with conceptual analysis of agent causation that many philosophers strongly feel.

One thing I should emphasize is that the subjectivity of agent causation in this sense doesn’t necessarily imply that there is no such thing as agent causation in the world. Consider again the case of colors. There is a clear sense in which the concept of red is subjective, but this doesn’t have to imply that there is no such property as red in the world. The property to which the concept of red correctly applies might well be a completely objective physical property. Similarly, I think, although there is something subjective about the concept of agent causation, what it applies to may be something completely objective and real.

7. Conclusion

In this chapter, we considered some attempts to conceptually reduce agent causation to event causation. Some extant proposals turned out to be defective. We also looked at some general concerns that many philosophers seem to share about the possibility of conceptual reduction, and tried to isolate the source of the problem. I think that it is because the concept of agent causation is essentially subjective that it is not conceptually reducible to event causation.

But all this is about the concept of agency and agent causation. Even if it is not definable in terms of event causation, the possibility still remains that what the concept of agent causation applies to may be facts about event causation. That is, even if conceptual
reduction is impossible, the possibility of *ontological* reduction seems to still remain. In the next chapter, I will try to make clearer a distinction between two kinds of reduction, conceptual and ontological, and then consider whether agent causation can be *ontologically* reducible to event causation.
CHAPTER 3

Agent Causation and Ontological Reduction

1. Introduction

In the last chapter, we saw that there is good reason to think that agent causation is not conceptually reducible to event causation. That is, statements of agent causation cannot be analyzed or translated into statements of event causation. Chisholm seems to have thought that this would show that we needed to accept agent causation as a *sui generis* kind of causation, insofar as there is agency. But on the face of it, this conclusion is not warranted. It is true that conceptual analysis will get us a kind of reduction, but it seems wrong to think that it is the only kind of reduction. On the contrary, it is well known that in many interesting cases of reduction in science and philosophy of mind, reduction is not a matter of conceptual analysis. Rather, it is a matter of showing that one kind of phenomenon is *ontologically* nothing “over and above” another kind of phenomenon. In short, conceptual reduction may be sufficient for reduction, but it is not a necessary condition.

Wasn’t Chisholm’s conclusion misguided by an obsolete conception of reduction? Shouldn’t he have considered the alternative kind of reduction? Perhaps agent causation is *ontologically*, but not conceptually, reducible to event causation, just as heat is generally thought to be ontologically reducible, though not conceptually reducible, to
more basic physical properties, like molecular kinetic energy. In this chapter, I want to explore this possibility. I have two specific aims in this chapter. First of all, I want to clarify what ontological reduction in general is supposed to be. I will claim that it has to conform to the so-called “functional model of reduction.” The second is to examine whether this reduction model can be coherently applied to agent causation. It is not my aim in this chapter to completely resolve this question. Rather, I want to get clearer on what should be the real issues in the debate about ontological reduction of agency and agent causation.

I will begin by making it clear that there is a distinction between conceptual reduction and ontological reduction. And I will consider the idea that what we have to seek as would-be reductionists about agent causation is ontological reduction rather than conceptual reduction. But for that, we need to look carefully at how ontological reduction in general is supposed to go. I will consider three different models of ontological reduction that are frequently discussed in the literature on reduction and reductionism, and claim that the functional reduction model is the most adequate one for ontological reduction. Once equipped with this model, I will consider whether this model can be applied to reducing agent causation to event causation. In particular, I will look at David Velleman’s approach to reduction of agent causation, which I believe is implicitly based on the idea of functional reduction. I will close the chapter by pointing to some remaining problems.
2. Conceptual reduction and Ontological reduction

What is it for one group of phenomena to be reduced to another apparently different group of phenomena? In the previous chapter, we took the question of reduction to be a question of translation or conceptual analysis, following Chisholm’s conception of reduction. Our concern was about whether the statement of agent causation can be analyzed or translated into the statement of event causation without any loss of meaning. But this way of understanding reduction is not the way in which many contemporary philosophers working in philosophy of science and philosophy of mind have conceived reduction, at least in recent decades.

To see this, consider a paradigm example of reduction in science: *heat to molecular kinetic energy*. What do we mean when we say that heat is reduced to molecular kinetic energy? Do we mean that we can express a statement about the former without loss of meaning into a set of statements about molecular kinetic energy? That doesn’t seem right. Indeed, many philosophers who are concerned with reduction and reductionism in recent debates explicitly deny that reduction is a matter of translation. They don’t think of reduction of heat to molecular kinetic energy as entailing that the concept or statements involving “heat” may be translated into those involving molecular kinetic energy.\(^{62}\) As a commentator on reduction and reductionism says:

> “Reductionism is no longer as the view that makes use of the logical positivist’s sense of reduction as translation. … Although one might allow that physical science contains many terms that are correlates of terms in the special sciences (to use Nagel’s example, ‘heat’ and ‘mean molecular

---

\(^{62}\) Logical positivists such as Rudolf Carnap think that reduction is a matter of translation. He says, “An object (or concept) is said to be *reducible* to one or more objects if all statements about it can be transformed into statements about these other objects” (Carnap 1928/1967, p. 6).
motion”), it is rarely supposed that these correlates are synonymous. Nagel, as discussed above, already noted this point. Even in the case where one might find two terms that refer to the same phenomenon, the terms themselves may differ somewhat in meaning, the identity of their referents being established empirically.”

Reduction is no longer understood as requiring that what is reduced (heat) be conceptually defined by its reducer (molecular kinetic energy). That is, even if what we mean by “heat” is not captured by the concept of molecular kinetic energy, it doesn’t prevent us from carrying out the reductive project.

Not only in philosophy of science, but also in philosophy of mind, philosophers have often been very explicit that the kind of reduction that they want is not conceptual reduction, but ontological reduction. For example, those who defend the identity theory of the mind, also called type physicalism—the view that mental properties are identical with physical properties—make the same point. For example, consider the following passages, the first by J. J. C. Smart and the second by Brian Loar:

“Let me first try to state more accurately the thesis that sensations are brain-processes. It is not the thesis that, for example “after-image” or “ache” means the same as “brain process of sort X”... It is that, in so far as “after-image” or “ache” is a report of a process, it is a report of a process that happens to be a brain process. It follows that the thesis does not claim that sensation statements can be translated into statements about brain processes.”

“We may take the phenomenological intuition at face value, accepting introspective concepts and their conceptual irreducibility, and at the same time take phenomenal qualities to be identical with physical-functional properties of the sort envisaged by contemporary brain science.”

63 Ney (2008), Sec. 2.
64 Smart (1959), p. 144.
65 Loar (1990), p. 598.
These philosophers make it clear that when it is claimed that mental properties are identical with physical properties, it does not necessarily mean that the two concepts have the same meaning, or that the statements involving the former can be translated into statements involving the latter. Rather, what they mean by the identity thesis may be that those phenomena which were previously assumed to be distinct are in fact one and the same phenomenon. As Loar puts it, although it may seem that the introspective concepts of mentality such as phenomenal qualities are irreducible, it may still be an open question whether those mental qualities are in fact identical with physical properties. That is, the kind of reduction they have in mind is “ontological reduction.”

From the foregoing discussion, I want to draw the following points: first, we can distinguish between two kinds of reduction, conceptual reduction and ontological reduction, and second, ontological reduction rather than conceptual reduction is what many philosophers are often concerned with in the debates involving reduction and reductionism, especially in philosophy of science and philosophy of mind.

Now, let’s go back to our case of agent causation. When we think of the reduction of agent causation to event causation, why shouldn’t we think of it as ontological reduction rather than conceptual reduction, as in the reduction of heat to molecular energy or sensations to brain processes? That is, isn’t it possible to ontologically reduce agent causation to event causation, even though it fails to be conceptually reducible—that is, the subjective concept of agent causation is never captured by the objective concept of event causation as we discussed in the Chapter II?

In fact, when some reductionists about agency maintain that the causal theory of action provides a reductive theory of agency, what they have in mind is ontological
reduction rather than conceptual one. There are several writers who have been very explicit on this. For example, Donald Davidson, after claiming that an intentional action is a behavior that is caused in the right way by attitudes that rationalize it, says this:

“We would not have shown how to define the concept of acting with an intention; the reduction is not definitional but ontological. But the ontological reduction, if it succeeds, is enough to answer many puzzles about the relation between the mind and the body, and to explain the possibility of autonomous action in a world of causality.”

Also, John Bishop, in his book *Natural Agency*, explores the possibility of ontological reduction of agent causation without conceptual reduction. He writes:

“Even if to act does not mean to be caused to behave by one’s reason, nevertheless, action can exist without the need for special agent causal relations in nature because actions are realized within complex sequences of events related only in event causal ways to one another. ... agent causation might be conceptually essential but not ontologically essential for action.”

Following these philosophers, Markus Schlosser also makes the same point.

“[The causal theory of action] does not claim to provide an analysis of our concept of agency in standard causal terms, but conditions for the realisation of agency by event-causal processes.”

Evidently, these philosophers make a distinction between the matter of defining or analyzing the concept of action in terms of event causation on the one hand, and the matter of considering the possibility of action’s being ontologically realized within the event causal order, on the other. They claim that what the causal theory of action purports to provide is the latter rather than the former; as Davidson says, the reduction is not

---

definitional (or conceptual), but ontological. That is, they seem to think that although agent causation may be conceptually distinctive and irreducible, it may be ontologically reducible to event causation.

But for the reductionist claim to be justified, it needs to be shown how the ontological reduction can be successfully achieved, rather than just asserting its possibility. To discuss this issue, I think we need to look closely at how ontological reduction is supposed to proceed in science and philosophy in general. This is what I will do in the next section. There are alternative models for ontological reduction, and I will claim that one model—the functional reduction model—is the most adequate model for ontological reduction. After that, I will examine whether this model can be applied to ontological reduction of agent causation.

3. Three Models of Ontological Reduction

In the current discussion of reduction and reductionism, three different models of ontological reduction have been influential among philosophers: the bridge-law reduction model, the identity reduction model, and the functional reduction model. In this section, I will review these models of reduction, and in the process, I will identify two requirements that ontological reduction has to meet. And then I will suggest that functional reduction is the only model that meets both requirements.

The bridge-law model was developed by Ernest Nagel in 1960s. His primary concern was with reduction of one theory to another theory, purportedly as a step toward theory unification. According to this model, a target theory is reduced to a base theory when the laws of the target theory are logically derived from the laws of the base theory
in conjunction with “bridge laws,” which connect the predicates of two theories. These bridge laws are typically construed as empirical and contingent law-like principles that connect between predicates or properties of the target theory and those of the base theory. These connecting laws play an essential role in the Nagel’s reduction, in that without them, the derivation from the base theory to the target theory cannot be achieved.

However, this requirement of bridge laws gives rise to a number of problems. But here let me focus on one problem that is directly relevant to our issue here. It is the difficulty that Nagel’s reduction based on bridge laws doesn’t result in ontological simplification. Jaegwon Kim raises the problem in the following passage:

“It is arguably analytic that reduction must simplify; after all, reduction must reduce. … On this score bridge laws of the form $M \leftrightarrow P$ apparently are wanting in various ways. Since $M \leftrightarrow P$ is supposed to be a contingent law, the concepts $M$ and $P$ remain distinct; hence bridge laws yield no conceptual simplification. Further, since we have only a contingent biconditional ‘iff’ connecting properties $M$ and $P$, $M$ and $P$ remain distinct properties and there is no ontological simplification.”

The source of the problem lies in the fact that bridge laws that serve as indispensable premises for derivation are taken to be contingent and empirical biconditional laws. The derivation with this form of bridge laws doesn’t genuinely reduce $M$ to $P$, ontologically or conceptually, because the two properties $M$ and $P$ in a bridge law still remain distinct. In other words, this model extends the reduction base by adding the bridge law premises involving the properties that are to be reduced. So, as Kim says, Nagel’s bridge-law

---

69 One main difficulty that philosophers generally raise, especially with regard to mind-body reduction, is that it is hard to obtain the bridge laws connecting between mental predicates or properties and physical predicates or properties. For example, “Davidson’s anomalist argument” and “the Putnam-Fodor multiple realization argument” are supposed to show that.

reduction “fails to give meaning to the intuitive ‘nothing over and above’ that we rightly associate with the idea of reduction.”\textsuperscript{71}

This flaw naturally leads some philosophers to require the bridge laws to take the form of identities, \( M = P \), instead of biconditionals, \( M \leftrightarrow P \). The idea is that if we enhance the bridge laws into identities, we will be able to get ontological simplification. That is, if \( M \) is identical with \( P \), there is a clear sense in which \( M \) is “nothing over and above” \( P \). This view is sometimes called “identity reduction.”\textsuperscript{72} On this view, as Kim describes, “reduction is accomplished by identifying phenomena and properties being reduced with appropriate items in the base domain.”\textsuperscript{73}

However, some argue that although identity reduction leads to ontological simplification, it doesn’t satisfy another crucial requirement for reduction. Take a familiar example. Suppose that empirical research shows that pain invariably correlates with C-fiber firing. But one may still feel that there remains a deep sense of mystery with such a correlation. Why does the feeling of pain arise, rather than, say, the feeling of itch, when there is C-fiber firing? Why does the feeling of pain arise from C-fiber firing rather than from some other neural states? As is often said, there is an “explanatory gap.”\textsuperscript{74}

We can easily see that Nagel’s bridge-law model fails to answer the question of the explanatory gap because what is to be explained is the bridge law itself—that is, why the correlation between pain and C-fiber firing holds—which is assumed as an unexplained

\textsuperscript{71} Kim (1998), p. 97.
\textsuperscript{72} This view traces back to the early mind-body identity theory endorsed by Feigl (1958) and Smart (1959). And a recent version of the view is defended by Hill (1991) and Block & Stalnaker (1999).
\textsuperscript{73} Kim (2007), p. 217.
\textsuperscript{74} For the problem of explanatory gap, see Levine (1983).
premise of the derivation in Nagel’s reduction. But if we upgrade the status of bridge laws from nomological correlations to identities, do we have a satisfactory answer to the explanatory question? In other words, does identifying pain with C-fiber firing close this gap?

It is hard to think so. After all, identities like “pain = C-fiber firing” is supposed to be a posteriori and have empirical content (although it is usually agreed, after Kripke, that it is a necessary truth). So far as it has empirical content, it must be something that can be justified and explained. And without an explanation, such an identity claim is bound to remain mysterious. As Kim says, “we must try to provide positive reasons for saying that things that appear to be distinct are in fact one and the same.”75 Unless we do this, there will remain a gap in our understanding. Thus, reductive identities appear to require explanatory support. We might call this the explanatory requirement on reduction.

Shouldn’t we then build this explanatory requirement into ontological reduction? There is a debate about whether meeting the explanatory requirement in general should be taken to be a necessary condition for ontological reduction. Suppose that there is very good reason to think that no explanation can be given for why pain is correlated with C-fiber firing. Would this vindicate that pain is not ontologically reducible to C-fiber firing? Some philosophers think so.76 But others deny that, saying that such an epistemological issue should be distinguished from the metaphysical issue.77 It is one thing that identity holds, and it is quite another matter whether we can explain it. They would admit that the

75 Kim (1998), p. 98. For this kind of objection to identity reduction, see Kim (2005), Chs. 4 and 5.
76 For example, Chalmers (1996) and Kim (2005).
77 For example, Hill (1991) and Block & Stalnaker (1999).
existence of such an explanation would indeed provide a very good ground to believe that pain is reduced to, or identical to, C-fiber firing. But the absence of an explanation (according to these philosophers) hardly shows that it is not reducible. For even without such an explanation, there may be other considerations, such as simplicity, that support the identity reduction.

We can’t go into this debate. But it seems to me that all parties should agree that explanatory reduction is always preferable, and that many cases of successful reduction yield reductive explanations. Moreover, unless we can provide such explanations, it will always be open to challenge the reductionist claim. And in any case, no harm will be done by working with the higher standard of reduction, as far as our aim is to endorse ontological reduction.

To sum up, there are two important conditions for a certain model of reduction to be an appropriate model for ontological reduction: “ontological simplification” and “explanatory requirement.” And the foregoing discussion shows that neither bridge-law reduction nor identity reduction is the best model for ontological reduction since the former doesn’t satisfy either requirement, and the latter doesn’t satisfy the second requirement, though it meets the first one. Now let us turn to the last model, that is, the functional model of reduction.

Functional reduction has been effectively advocated by many philosophers recently as an alternative model over bridge-law reduction and identity reduction, purporting to supplant, and improve on, both models. Kim is one of the representative philosophers who criticize both Nagel’s bridge-law reduction and identity reduction for the reasons above, and adopt the functional model of reduction as an appropriate model for mind-
body reduction. He describes the core idea of the functional model of reduction as follows:

“To reduce a property, say being a gene, on this model, we must first ‘functionalize’ it; that is, we must define, or redefine, it in terms of the causal task the property is to perform. Thus being a gene may be defined as being a mechanism that encodes and transmits genetic information. That is the first step. Next, we must find the ‘realizers’ of the functionally defined property—that is, properties in the reduction base domain that perform the specified causal task. It turns out that DNA molecules are the mechanisms that perform the task of coding and transmitting genetic information—at least, in terrestrial organisms.”  

Joseph Levine takes a similar view:

“Stage 1 involves the (relatively? or quasi?) a priori process of working the concept of the property to be reduced “into shape” for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Step 2 involves the empirical work of discovering just what those underlying mechanisms are.”

Functional reduction consists of two steps. The first step is a conceptual or _a priori_ process while the second step is an empirical or _a posteriori_ process. What we need to do first is to define or redefine the concept of the target property in terms of its function—that is, its causal role. So this process is grounded in conceptual analysis. But the conceptual analysis in functional terms alone can’t get us reduction. What is to be done next is to empirically discover the physical realizer or mechanism that in fact plays the role definitive of the property to be reduced. This process is _a posteriori_, which can be

---

80 But Kim points out that step (1) is a conceptual process, but not necessarily an _a priori_ process. He says, “how a functional definition for a given M is to be formulated is likely to have to be informed by empirical findings and theoretical needs” (Kim 2007, p. 222).
accomplished through empirical research. Then, it is supposed to follow from those two steps that the target phenomenon is identical with the realizer in the base domain.

Let me try to make this little more precise. Let M be a property that is the target of reduction, and P be a property or mechanism in the base domain. Then we can formulate the process of functional reduction as follows:

1. For x to have M, by definition, is for x to be in a state that plays causal role R.
2. P (a property instantiated by, or in, x) is the mechanism that plays causal role R.
3. Therefore, for x to have M = for x to have P.

For example, the functional reduction of pain may go as follows:

4. For x to be in pain, by definition, is for x to be in a state that is apt to be caused by tissue damage and apt to cause wincers, groans, etc.
5. C-fiber firing (instantiated in x) is the mechanism that is apt to be caused by tissue damage and apt to cause wincers, groans, etc.
6. Therefore, for x to be in pain = for x to be in the state of C-fiber firing.

Here (4) specifies the functional role of pain, and this is supposed to be a conceptual matter. (5) states that as a matter of fact, what realizes this role is C-fiber firing. (6) concludes from these that the particular instance of pain is identical to the instance of C-fiber firing.

Now we need to show that this model of reduction satisfies the two requirements of ontological reduction. Does functional reduction yield ontological simplification? I think that there should be no doubt that (4) and (5) together give us the identification of x’s
having pain with x’s having C-fiber firing—that is, *token identity* between a particular instance of pain and a particular instance of physical states. But it doesn’t guarantee that pain itself is identical to C-fiber firing. And as a matter of fact, this may well be false; for the physical state that realizes the causal role may well be different across species (or even different across individual human beings); that is, as is often said, pain is “multiply realizable.” There has been a debate on the ontological implication of this, which we can’t discuss here. But many philosophers seem to agree that token identity is enough for ontological simplification.\(^81\)

Secondly, does functional reduction satisfy the explanatory requirement? Recall that we identified this with the demand that “we must try to provide positive reasons for saying that things that appear to be distinct are in fact one and the same.”\(^82\) (4) and (5) together (if they are true) can provide compelling positive reasons to believe that x’s being in pain is identical to x’s being in the state of C-fiber firing. Moreover, we can adequately respond to the explanatory demand such as “Why did x have pain, rather than itch, when x was in the state of C-fiber firing?” in the same manner.\(^83\) So functional reduction seems to satisfy the explanatory requirement.

Now, for our purpose, it is important to note that in general, there is no guarantee that functional reduction of certain phenomena relative to some selective base domain will succeed. It can fail for at least two different reasons, corresponding to the two stages

\(^{81}\) For a classic paper that discusses this issue, see Kim (1992). For a recent discussion, see Kim (2007).

\(^{82}\) Kim (1998), p. 98.

\(^{83}\) For a detailed formulation of how the functional reduction responds to the explanatory demand such as “why did x have pain when x was in the state of C-fiber firing?”, see Kim (2007), p. 223.
of functional reduction. First of all, for some reason, it may not be possible to complete the first step; that is, certain phenomena may fail to be functionally analyzable. In fact, some dualists believe that pain and other experiential states are such phenomena. They say that any functional analysis of pain in terms of its causal roles is bound to leave out what is essential to pain; that is, a special kind of qualitative feeling. So, according to these philosophers, pain is not functionally (hence, ontologically) reducible to a physical property, and therefore the explanatory gap cannot be closed.

There can be a different reason why functional reduction fails. In step 2 of functional reduction, we are supposed to find the realizer of the functional role defined in step 1 in the reduction base. In general there is no guarantee that we can succeed in finding such a realizer in the reduction base. However, even if we haven’t found a realizer yet, the target phenomenon will remain reducible, although not actually reduced, to a phenomenon in the base domain. For whether the second step of functional reduction will succeed is largely an empirical matter. So philosophically, the more important step of functional reduction will be step 1. This point will be important to our discussion of the reduction of agent causation.

Now we will return to the case of agent causation. In what follows, I will investigate whether the functional model can be applied to the reduction of agent causation to event causation. In particular, I want to consider a proposal made by David Velleman, who more or less explicitly applies the functional model of reduction to agent causation.

---

84 For the distinction between reduction and reducibility, see Kim (2007).
4. Velleman’s Reductionist Project and Functional Reduction

Given that the functional reduction model has been influential in discussion of reduction (especially reduction of the mental to the physical), it is a bit strange that reduction of agency and agent causation has been rarely discussed within a similar framework. But Velleman is a notable exception. In his influential essay “What Happens When Someone Acts?”, he tries to reduce agent causation to event causation through something very much like functional reduction. In this section, I will consider his view, and try to make more explicit how the functional reduction model can be plausibly extended to agent causation.

Velleman is one of the philosophers who think that the notion of agent causation has to be central in the discussion of agency. According to him, “our concept of full-blooded action requires some event or state of affairs that owes its occurrence to an agent and hence has an explanation that traces back to him.”85 He sees the problem of agency as that of reconciling this commonsense conception of agency that implicates agent causation with our scientific view of the world, which “regards all events and states of affairs as caused, and hence explained by other events and states or by nothing at all.”86

And his way of doing this is to reduce agent causation to event causation. He says:

“Such a reconciliation will have to show how the causal role assigned to the agent by common sense reduces to, or supervenes on, causal relations among events and states of affairs. And the agent’s being a supervenient

cause of this sort might also be called agent-causation, in a more relaxed sense of the phrase." 87

Here Velleman seems to imply that reduction of agent causation to causal relations among events is similar in relevant respects to reduction of other phenomena. It is worth noting that he explicitly likens the problem of agency conceived in this way to the mind-body problem.

“The problem of agency is … independent of, though indeed parallel to, the mind-body problem. Just as the mind-body problem is that of finding a mind at work amid the workings of the body, so the problem of agency is that of finding an agent at work amid the workings of the mind.” 88

The mind-body problem, as Velleman conceives it, is the problem of “finding a mind at work amid workings of the body.” Many contemporary philosophers think that something like functional reduction of the mental properties to the physical properties, if successful, could solve this problem. So if the parallel between the mind-body problem and the problem of agency holds, we may expect that something like functional reduction of agent causation to “the workings of the mind” will solve the problem of agency.

So specifically how is such a reconciliation or reduction supposed to go? The following passage contains the core idea:

“The way to advance the reductionist project is … to get the process of reduction going, by breaking agent-causation into its components. And surely, the principal component of agent-causation is the agent himself. … then, we should look for events and states to play the role of the agent. … I pointed out that the agent intermediates in various ways between his reasons and intentions, or between his intentions and bodily movements …What plays the agent’s role in a reductionist account of agent-causation will of course be events or states—most likely, events or states in the

88 Velleman (1992), P.131.
agent’s mind. We must therefore look for mental events and states that are functionally identical to the agent, in the sense that they play the causal role that ordinary parlance attributes to him.”

Although he doesn’t explicitly appeal to any specific model of reduction, what Velleman has in mind here looks exactly like functional reduction of agent causation. What we have to do to reduce agent causation is, first, to find “the causal role that ordinary parlance attributes to the agent,” and then “look for mental events and states” that play the role. In short, as Velleman says, “a reductionist philosophy of action must … locate a system of mental events and states that perform the functional role definitive of an agent.”

But there is a potentially significant difference between other familiar cases of functional reduction and the case of reduction of agent causation conceived by Velleman. He says that we first need to “[break] agent causation into its components,” the principal component of which is “the agent himself.” What is the point of the inclusion of this additional step? I think that Velleman, in effect, is responding to a potential worry that may be raised against applying the functional reduction model to agent causation. Let me explain what this worry is, and how “breaking agent causation into its components” can dissolve it.

As we saw, what is fundamentally important in functional reduction is to functionally analyze or define the target of reduction. What is it to functionally define something? In familiar cases of reduction, such as pain to C-fiber firing or heat to molecular kinetic energy, we may safely assume the following:

---

To functionally define F is to define it by specifying the causal role of F.

For example, to functionally define pain is to define it by specifying its causal role. And it is precisely because pain is assumed to be not functionally definable in this way that it is thought to be irreducible to states in the base domain. But then can we really functionally define agent causation in this sense? If we apply (*) directly to agent causation, what we get is the following: to functionally define agent causation is to define agent causation by specifying the causal role of agent causation. But this doesn’t seem to make much sense. What could it mean to specify the causal role of agent causation? So one might think that agent causation, unlike other familiar objects of reduction, is not really a kind of thing that can be functionally defined, and hence not a kind of thing to be functionally reduced to something in the base domain.

However, we can easily get around this apparent difficulty, by, as Velleman says, “breaking agent causation into its components.” Obviously, what has the causal role in an agent’s causing an action is not agent causation itself, but the agent. Then we may be able to define agent causation by specifying the causal role of the agent in instances of agent causation. That is, the following may be conceptually true (for some appropriate R): S’s causing action A = S’s playing causal role R in producing A.

This seems to be enough to get the functional reduction process going. We can formulate the process of reduction of agent causation Velleman has in mind as follows:

1) Agent S’s causing action A = def S’s playing role R in producing A.

2) S’s mental states M play role R in the production of A.

3) Therefore, S’s causing A = M’s causing A.

63
This schema is little different from the one we considered in the last section. But it does contain all essential elements of functional reduction. (1), again, is supposed to be a conceptual matter, and (2) is expected to be filled in by empirical research. The conclusion (3) seems to follow from (1) and (2). So, if this works, we will get explanatory reduction of agent causation.

As we saw, philosophically, (1) is the critical step. Velleman is explicit that it is a conceptual task: It is supposed to be “assigned … by common sense” or attributed by “ordinary parlance.”\(^91\) He says:

“We say that the agent turns his thoughts to the various motives that give him reason to act … We say that the agent calculates the relative strengths of the reasons before him … We say that the agent throws his weight behind the motives that provide the strongest reasons …”\(^92\)

What matters is what we ordinarily think and judge about the agent’s role in producing actions. The role our common sense attributes to the agent, according to Velleman, is to consider various reasons, and calculate strengths of them, and “throw his weight behind the motives.” He is also aided by our intuitive judgments about particular cases of actions (recall the cases of defective actions we considered in the last chapter), which is also assumed to be a conceptual matter. Eventually, he proposes the following:

“In a full-blooded action, an intention is formed by the agent himself, not by his reasons for acting. Reasons affect his intention by influencing him to form it, but they thus affect his intention by affecting him first. And the agent then moves his limbs in execution of his intention; his intention doesn’t move his limbs by itself. The agent thus has at least two roles to

\(^91\) Velleman (1992), p. 130 and p. 137.
\(^92\) Velleman (1992), p. 141.
play: he forms an intention under the influence of reasons for acting, and he produces behavior pursuant to that intention.”

When an agent’s action is directly caused by his reasons (beliefs and desires), the agent himself is not participating in the action. The agent’s role is supposed to be “to intervene between reasons and intention, and between intention and bodily movements, in each case guided by the one to produce the other.” It is only when the agent plays such a role that the agent himself can be said to cause the action.

So that is the proposal made by Velleman as for step 1 of functional reduction of agent causation. Is he right about it? I think that his account must be on a right tack (especially about the role of “intervening between reasons and intention”). However, I think that there is much more that needs to be said in order for the functional reduction to work. My main concern is this. Velleman says that “reasons affect [an agent’s] intention by influencing him to form it.” But precisely how do reasons “influence” the agent? Is it some sort of causal relation, or something different? It seems to me that insofar as the agent’s role is located in relation to reasons for actions, we first need to get clearer on the roles that reasons play in producing actions, in order to correctly identify the agent’s role in it. I will pursue this issue in a later chapter, but for a moment, let us suppose that something like this is a correct analysis of the agent’s role in producing actions.

Let’s turn to the second step of functional reduction. Velleman says:

“Of course, the agent’s performance of these roles probably consists in the occurrence of psychological states and events within him. To insist that the story mention only the agent himself as the object of rational influence, or as the author and executor of intentions, would be to assume a priori

94 Velleman (1992), P. 125.
that there is no psychological reduction of what happens in rational action. One is surely entitled to hypothesize, on the contrary, that there are mental states and events within an agent whose causal interactions constitute his being influenced by reason, or his forming and conforming to an intention.\footnote{Velleman (1992), P. 124.}

In the end, Velleman has a specific proposal about what realizes the agent’s role (which we will consider in a later chapter). But for our present purpose, what is important is that we are “surely entitled to hypothesize” that there are mental events or states that realize this role. As in functional reduction in other cases, it is supposed to be \textit{a posteriori} and empirical. At present, we may not have the knowledge of what realizes the agent’s role in particular cases, and so perhaps we don’t have the actual \textit{reduction} of agent causation. But as we saw in the previous section, this doesn’t affect \textit{reducibility} of agent causation.

If all this is successful, we will come to have functional reduction of agent causation. As we emphasized in the last section, the functional reduction model is supposed to yield \textit{ontological simplification}. The same is true of the functional reduction of agent causation. It gives us at least the identification of a \textit{particular instance} of agent causation with an instance of event causation—token identity. As before, agent causation may be realized by different mechanisms in different instances. But as I said before, token identity is enough for ontological simplification.

The functional reduction of agent causation will also meet the \textit{explanatory requirement}, in that it provides positive reasons to identify an instance of agent causation with an instance of event causation. Those who think that agent causation must be something over and above event causation tend to express their sense of mystery, asking “How can the agent’s mental states add up to the agent’s activity?” Now a reductionist

\footnote{Velleman (1992), P. 124.}
can give the following answer: “The agent’s activity, definitionally, consists in the agent’s playing role R in producing the action, but R is played by some mental states and events. And that’s how the agent’s mental states add up to the agent’s activity.”

So far, I don’t see any obstacle to extending the functional reduction model to agent causation, although I do think that the critical step of functional reduction of agent causation, that is, the task of specifying the agent’s role in producing actions, remains to be worked out further. This I will take up in the remaining chapters. But there is a more general worry that may be raised against the whole project of functional reduction of agent causation. Before closing this chapter, I want to consider this worry.

5. Distinctness of Levels

Take again the case of the mental-physical reduction. I think that the picture that reductionists have in mind is something like this. We can describe a person’s state by saying that the person is in pain, and also by saying that the person is in the state of C-fiber firing. But we are actually describing one and the same phenomenon on two different “levels of description.” The fact that they describe ontologically the same phenomenon cannot be conceptually obvious, however, because they belong to disparate sets of vocabulary; after all, one can master the mental vocabulary without mastering the physiological vocabulary. Some principle is required bridging these distinct levels, and this is supposed to be provided by the conceptual analysis of pain and empirical work of identifying the realizer of its role (which amount to the two stages of the functional reduction.)

67
Now apply this picture to ontological reduction of agent causation. The reductionist will want to say exactly the same in this case. We describe or explain an agent’s action by saying that it is caused by the agent, and also by saying that it is caused by the agent’s mental states and events. But we are actually describing one and the same phenomenon in *two distinct levels of description*. This cannot be conceptually obvious (and thus agent causation is conceptually irreducible), because they belong to different levels of description. Perhaps one can master one set of vocabulary without mastering the other. Some principle is required bridging the two distinct domains, and this principle is supposed to be provided by the functional analysis of an agent and empirical work of identifying the realizer of the role.

As it is clear from these descriptions, the whole approach of ontological reduction seems to presuppose that the phenomenon to be reduced belongs to a different *level* or *domain* from the phenomenon that does the reducing.\(^96\)

Now a critic may ask: This talk of different levels seems unproblematic in the case of pain and C-fiber firing. But is it right in the case of agent causation? That is, when we say that an agent brings about an action, and when we say that mental events and states cause an action, are we employing two different levels of description? Don’t they both belong to *one and the same level or domain*—both to the mental domain? The functional reduction model presupposes that what is reduced and what does reducing belong to distinct domains. But then this model would not be applicable to reduction of agent

---

\(^{96}\) Kim says: “the idea of reduction, and reductive explanation, involves two distinct domains or levels. ... a reductive explanation of a phenomenon must be expected to do its job by resorting only to explanatory resources at a different level, a “lower” level in relation to the level of the phenomenon to be explained” (Kim 2005, p. 106).
causation to event causation, because they belong to the same level. Worse, ontological reduction without conceptual reduction seems to make sense only when what is reduced and what does reducing belong to distinct levels. If this is right, then the only kind of reduction that can reduce agent causation to causation by mental states and events should be conceptual reduction. So after all, Chisholm had a point when he said that if agent causation is not conceptually reducible to event causation, then we have to accept *sui generis* agent causation!

How should the proponent of ontological reduction respond to this sort of challenge? I think that one has to agree that the applicability of the functional reduction model, and indeed ontological reduction in general, depends on the distinctness of levels. But I think the critic is hasty when he said that there is no distinctness in levels involved in the case of agent causation, just because they both belong to “the mental domain.” I think that those who seek ontological reduction of agency and agent causation implicitly assume that agent causation and causation by mental events and states belong to different levels. For example, as we saw, when Velleman says that an agent as a cause may be “supervenient cause,” or likens the issue to the mind-body problem, he was implicitly supposing that agent causation and causation by mental events and states belong to the distinct levels. But they are rarely explicit what these levels could be.

I believe that there is a good sense in which the levels of description are indeed different. This is in fact related to the point I made in the earlier chapter. Agent causation is something visible only from the first-person point of view, and there seems to be something essentially *subjective* about the concept of agent causation. On the other hand, there is nothing particularly first-personal or subjective with describing or explaining an
action in terms of causation by mental states and events. If this is right, then it seems not unreasonable to think that there are distinct levels of description involved in agent causation and event causation. Of course, a lot more needs to be said about this distinction, and this is what I will try to do in the next chapter. I think that once we make this distinction clearer, the worry we’ve just raised can be dispelled. Moreover, I think it will become clearer what the role of an agent is supposed to be in producing actions when it is seen from the first-person point of view.

6. Conclusion

Many philosophers of action with a reductionist inclination don’t seem to be impressed by the fact that agent causation cannot be conceptually reduced to event causation. It is because they think that reduction proper should be ontological rather than conceptual. But I think that they pay too little attention to the nature of ontological reduction. So that’s what provoked the present chapter. I tried to make it clear what ontological reduction is supposed to be, and explored the possibility of applying the functional reduction model to agent causation.

As it turned out, there is no serious obstacle to pursuing this project. But we identified two tasks that need to be done. One is to make more precise what role an agent plays in producing actions. The other is to get clearer on distinct descriptive levels to which agent causation and event causation belong. I will address these problems in the concluding chapter. But for that, I first need to say something about action explanation and different “perspectives” on action.
CHAPTER 4

The Internal Perspective, Action Explanation, and
Agent Causation

1. Introduction

In Chapter II, we examined what reasons can be provided for conceptual irreducibility of agent causation to event causation. In the end, I suggested that agent causation is not conceptually analyzable into event causation because they belong to disparate conceptual schemes. Agent causation is something discernible only from the first-person point of view, and it is “essentially subjective,” while event causation is completely objective. The same idea played a role in Chapter III too. I explored the possibility of applying the functional reduction model to agent causation. But one worry that emerged toward the end of the chapter was that the whole idea of ontological reduction seems to presuppose a system of distinct “levels.” And I finished the chapter by suggesting that different perspectives may constitute different levels of description.

In this chapter, I want to take up this issue of different “perspectives” on action directly and in isolation. Thomas Nagel is perhaps the single most important philosopher in this regard. In his *The View from Nowhere*, he explicitly distinguishes two different
points of view on actions—internal and external, or subjective and objective—and claims that agency, freedom, and responsibility and other related phenomena are only visible from the internal point of view.97

I want to take Nagel’s own thoughts on this matter as a reference point. He thinks that the internal perspective on our actions is essentially tied to a peculiar way of explaining actions. We explain our actions through “intentional explanation,” by referring to justifying reasons, and such explanations can’t be regarded as causal explanations that are formulated from the objective perspective. Drawing on Nagel’s insight, I will claim that this type of explanatory practice is constitutive of the internal perspective on action. Nagel in the end thinks that there is something unintelligible about the internal perspective of action, because intentional explanations of actions aren’t really adequate as explanations. But I will argue that we can make good sense of such explanations. And the discussion will naturally bring us back to agent causation and its reducibility.

Here is my plan. In Section 2, I will start with Nagel’s discussion of the internal perspective on actions, and raise some questions. In the next three sections, I will try to answer those questions. First, I will examine the nature of what he calls “intentional explanation,” drawing largely on Jaegwon Kim’s work on the topic. And then I will consider how intentional explanation is related to the internal perspective on actions. Lastly, I will examine Nagel’s concern about unintelligibility of intentional explanation. I will finish the chapter discussing the “explanatory exclusion” problem applied to intentional explanation.

97 Nagel (1986), Chapter 7.
2. Two Perspectives on Action

Nagel says:

“From the inside, when we act, alternative possibilities seem to lie open before us: to turn right or left, to order this dish or that, to vote for one candidate or the other—and one of the possibilities is made actual by what we do. The same applies to our internal consideration of the actions of others. ... From an external perspective ... the agent and everything about him seems to be swallowed up by the circumstances of action; nothing of him is left to intervene in those circumstances. This happens whether or not the relation between action and its antecedent conditions is conceived as deterministic. In either case we cease to face the world and instead become parts of it; we and our lives are seen as products and manifestations of the world as a whole. Everything I do or that anyone else does is part of a larger course of events that no one “does,” but that happens, with or without explanation.”

When I consider my own action “from the inside,” I find it almost impossible to think of it as completely determined and explainable by antecedent events and states—including my mental events and states. I take my own action to be “up to me.” But when I consider other people’s actions from an external perspective, I regard them as determined and explainable by their beliefs, motives and other mental states and circumstances. The agency is “swallowed up by the circumstances of action.”

As Nagel notices, it is true that we sometimes take an internal point of view about other people’s actions too, sympathetically putting ourselves in other people’s shoes. It is also true that we sometimes take an external perspective toward our own (especially past) actions (although, as many philosophers observed, it seems impossible that one takes the

---

external point of view at the moment of decision\textsuperscript{99}). All the same, the two perspectives seem to be sharply distinguished.

According to Nagel, agency, freedom, autonomy, and responsibility are visible only from the internal point of view. (We may include in the list \textit{agent causation}). So understanding the internal perspective will be central to understanding them.

I find the idea that there are two quite different “perspectives” or “points of view” on action not only intuitively compelling, but also theoretically very attractive. We have been struggling to reconcile two apparently conflicting conceptions of actions. One is that our actions are causally determined by antecedent events including our beliefs and desires, and the other is that we ourselves, not states or events that happen in us, bring about our actions. The reason why such reconciliation looks difficult, if not utterly impossible, is that they appear to be mutually exclusionary. If I think that I brought about my action, but you think that it is caused by antecedent events such as my beliefs and desires, then aren’t we \textit{contradicting} each other, so that at best one of us can be right? But the talk of “perspectives” or “points of view” suggests a different kind of relationship between the two. It will help to take an analogy. Suppose that you and I are seeing a table from two different positions, or \textit{from two different perspectives}. We will experience the table differently. But can we say that our experiences are “contradictory,” so that only one of us is representing the table correctly? Intuitively, this doesn’t seem right. A correct thing to say seems to be that both you and I are representing the table correctly, but from different perspectives. Then, applying a similar thought, when you consider my action as

\textsuperscript{99} For example, see Ginet (1962).
determined by events, and I consider it as determined by myself, we may not be contradicting each other; we both are viewing it correctly, *from different perspectives*.

However, of course, we can’t rely entirely on an analogy to make such a substantial point. We should be able to say something about what these different points of view or perspectives consist in, hopefully in the way that vindicates the initial attractiveness of the notion of a perspective.\(^{100}\) And that seems to pose a deep philosophical challenge.

Nagel expresses skepticism as to the possibility of giving a coherent account of the internal perspective on action. He says:

> “The difficulty … is that while we can easily evoke disturbing effects by taking up an external view of our own actions and the actions of others, it is impossible to give a coherent account of the internal view of action which is under threat. When we try to explain what we believe which seems to be undermined by a conception of actions as events in the world … we end up with something that is either incomprehensible or clearly inadequate.”\(^{101}\)

But why? Why is he so skeptical about giving an account of the internal perspective on action? It is related to difficulties he sees concerning what he calls “intentional explanation,” in which actions are “fully explained only intentionally, in terms of justifying reasons and purposes.”\(^{102}\) He says:

> “The objective view seems to wipe out such autonomy because it admits only one kind of explanation of why something happened—causal explanation—and equates its absence with the absence of any explanation at all. … There is no room in an objective picture of the world for a type of explanation of action that is not causal. The defense of freedom requires

\(^{100}\) But some philosophers may think that we can leave the notion of the perspectives as primitive. My impression is that Kantians often take such a position. See, e.g., Korsgaard (1989).

\(^{101}\) Nagel (1986), pp.112-3.

the acknowledgment of a different kind of explanation essentially connected to the agent’s point of view.”

The internal perspective on action is supposed to be tied to the kind of explanation that is “not causal” and is “essentially connected to the agent’s point of view”; that is, “intentional explanation.” So making sense of agency, which is only visible from the internal perspective, requires us to make sense of intentional explanation. But Nagel claims that such explanation is unintelligible, saying that “the alternative form of explanation doesn’t really explain the action at all.” (We will consider his argument for this more closely in a later section.)

There are many questions to be asked here. Let me divide them into three groups:

(I) What is an “intentional explanation,” and how do “justifying reasons” explain actions? Why are such explanations “comprehensible only through my point of view”? And how is it different from what Nagel calls “causal explanation” that is offered from the external standpoint?

(II) What does the internal perspective have to do with what he calls “intentional explanations”? Why is making sense of intentional explanations required to make sense of the internal view of action?

(III) Why does Nagel think that such explanation is unintelligible? Is he right that an intentional explanation of an action “doesn’t really explain the action at all”?

I will deal with these three groups of questions in the next three sections, in that order. My ultimate purpose is to give a coherent and plausible account of what the internal perspective on action consists in. I’ll endorse, and develop, Nagel’s insight

---

contained in (I) and (II). I believe that Nagel is right about essential subjectivity of “intentional explanation.” And I agree with him that the internal perspective is intimately connected to the practice of our explaining our own actions. But as for (III), I want to oppose Nagel, and claim that there is nothing incoherent about such explanations. The discussion will naturally lead us back to the issue of agent causation and its reducibility.

3. Action Explanation from the Agent’s Point of View

Nagel speaks of a “kind of explanation connected to the agent’s point of view.” Explanations of this kind are supposed to explain actions “intentionally, in terms of justifying reasons and purposes.” This kind of explanation is contrasted with what he calls “causal explanation,” which is tied to the external point of view. In this section, I want to consider the nature of such explanations. I will draw largely on Jaegwon Kim’s recent paper “Taking the Agent’s Point of View Seriously in Action Explanation.”

Let us start with the kind of explanation with which “intentional explanation” is contrasted. I think we can safely assume that the standard view of explanation is largely the Hempelian one. According to this conception, an explanation takes the form of an argument, whose conclusion is a statement describing the phenomenon to be explained, or the explanandum. At least one of the premises, which collectively constitute the explanans, should be an empirical law. The laws involved may be universal laws (mainly in hard sciences), or statistical or ceteris paribus laws (mainly in special sciences). These laws, taken together with the singular premises stating initial conditions, must imply the explanandum, either deductively, or inductively. To emphasize the involvement of laws,
let us call this sort of explanation *nomological explanation*. (And we may take what Nagel means by causal explanation to be such explanation.)

There is a debate about exactly what makes explanations explanatory. There seems to be a general consensus that laws are invariably involved in explanations, but philosophers disagree about the reason why laws must be involved. Hempel himself seems to put emphasis on the epistemological side, and claims that what is central is that an explanation “affords good grounds to believe that the phenomenon to be explained did, or does, indeed occur.”¹⁰⁵ In other words, what makes explanations explanatory is “nomic expectability.” But other philosophers think that what is essential is *causal relation* between an explanans and its explanandum. That is, it is not the presence of laws *per se*, but causal relations that are usually expressed by laws, that does the explanatory work. For our purpose, we don’t have to choose one of these views; it is enough to say that both have some plausibility.

This model of nomological explanation is supposed to be universal, in that it can be applied to explanations of any phenomena, including *human actions*. According to the Hempelian conception of explanation, there is nothing special about explanations of actions. There are supposed to be many “laws” concerning human actions, perhaps some folk-psychological, but others more scientific. And we can explain actions by subsuming them under those laws. In particular, explanations of actions in terms of the agent's beliefs and desires can be rendered as nomological explanations in this sense. That is, they fit the following general schema:

(A) S desired that p, and believed that doing X would bring it about that p.

Rational agents who desire that p and believe that doing X will bring it about that p tend to do X.

Therefore, S did X.

Isn’t this adequate as a model of explaining and understanding our actions?

Kim’s complaint with the adequacy of this as a model for explaining and understanding our own actions is well summarized in the following passage:

“To understand what I did, I have to know what all agents like me would do in similar circumstances. Prima facie, that seems impractical and unrealistic, if not outright impossible. Moreover, it just doesn’t seem the way an agent’s self-understanding works—having that kind of nomological information seems neither necessary nor sufficient—indeed it seems entirely irrelevant—for my understanding of my own actions.”  

As we saw, what is essential in nomological explanation is the inclusion of empirical laws. So in order for the agent to explain his own action nomologically, he has to know relevant laws that subsume his action. But Kim says that having that kind of information is “impractical and unrealistic,” and even “irrelevant.” One might protest that it is not so impractical and unrealistic, because folk-psychological laws are a priori. This may make nomological explanation of my own actions less practical and unrealistic, but the point of irrelevance still remains. That all rational agents tend to act in a certain manner seems completely irrelevant to understanding my own actions.

Then how do we explain and understand our own actions? Here is the basic principle that Kim finds “self-evident and wholly compelling”:

“To know, or be able to reconstruct, the practical deliberation that led me
to decide to do something (and actually do it) is a fully satisfactory way of
understanding why I decided, and acted, as I did.”

Typically, we may assume, when an agent performs a certain intentional action, she
engages in some sort of practical deliberation or practical inference, either consciously
or unconsciously. And simply by “retracing the process of deliberation,” the principle
says, we get explanations of our own actions. Let us call this Kim’s principle.

In order to understand how such explanations actually work, we first need to see
how practical deliberation or practical inference is supposed to go. Kim offers the
following schema:

(B) I am in circumstances of kind C.

In circumstances of kind C, the appropriate thing to do is X.

So I ought to do X.

Notice that the conclusion of this reasoning is an “ought” statement. (So this is not yet an
explanation of why I did X.) Here is an example from Kim:

(C) It is going to rain this afternoon.

Unless I have an umbrella with me, I am going to get wet.

I don’t want to get wet.

In this situation, the right thing to do is to take an umbrella.

Therefore, I ought to take an umbrella.

Let us consider each component of the practical inference in detail. I take the first
premise of schema (B) (“I am in circumstances of kind C”) to be the placeholder for the
considerations that the deliberator takes to justify his doing X, that is, justifying reasons.

Kim (2010a), p. 139.
So in example (C), that it is going to rain, that I don’t want to get wet, etc, constitute my reasons for taking an umbrella. As Kim emphasizes, they are statements that *I believe to be true*, but typically, they will not be statements *about my belief*—for example, the statement that *I believe* that it is going to rain this afternoon.\(^{108}\) That is, the reason that justifies my taking an umbrella is not something about my state of mind, but something about the weather. (But in view of this, it may seem a bit awkward to include a statement about my *desire*, as Kim does in (D). Shouldn’t it be rather something about getting wet, such as the *undesirability* of getting wet, rather than my attitude toward it?)\(^{109}\) Some philosophers may take issue with this, and insist that justifying reasons are my mental states. But this won’t affect our discussion. What is essential is that they are reasons that *justify* an action for the deliberator.

Turn to the second premise of (B), “In circumstances of kind C, the appropriate thing to do is X.” I think it is very important to properly understand what this premise does. In our actual deliberation, such a judgment may not appear explicitly. (“I ought to take an umbrella, because it is going to rain this afternoon.”) Nonetheless, such a judgment should be in the background, in order for practical inferences to work. Considerations such as that it is going to rain this afternoon, etc, become justifying reasons for me only if I *take* them to be reasons for taking an umbrella with me. That is, if I did not *judge* the proposition that it is going to rain this afternoon to be a reason for taking an umbrella, I wouldn’t judge that I ought to take an umbrella, even if I believe

\(^{108}\) Kim (2010a), p. 139.  
\(^{109}\) T. M. Scanlon forcefully argues that justifying reasons are propositions believed, rather than mental states. And he also claims that typically, justifying reasons don’t need to include desires. See Scanlon (1998), Sec. 1.9.
that it is going to rain this afternoon. Practical reasoning is always mediated by such
normative judgments, implicitly or explicitly. I take it that the second premise of (B)
expresses such a judgment. Even those who think that justifying reasons are mental states,
rather than propositions believed, should not deny this point.

Assuming that this is a plausible picture of how practical deliberation goes, let us
see how Kim’s principle applies. According to Kim’s principle, simply by recapitulating
practical deliberation leading to doing X, we get an explanation of why I did X. So we
have the following schema:

(D) I was in circumstances of kind C.

The appropriate thing to do in C was X.

Therefore, I did X.

Notice that the only difference between this and (B) is that “I did X” replaces “I ought to
do X.” Using our example, (C) can be transformed into an explanation of my action.

(E) It was going to rain this afternoon.

Unless I had an umbrella with me, I was going to get wet.

I didn’t want to get wet.

In this situation, the right thing to do was to take an umbrella.

Therefore, I took an umbrella.

So assuming that Kim’s principle is true, it simply follows that (E) constitutes an
adequate explanation by which I understand my own action. And this seems intuitively
right. Let us call this sort of explanation of action normative explanation, to emphasize
the role of normative judgments in it.
It will help to compare (A) (that is, a nomological explanation of an action in terms of beliefs and desires) with (D). They both explain an action by referring to what may be properly called “reasons” for actions. But they play different roles. In the case of nomological explanation, what mediates between reasons and my doing X is a certain general statement, or a law. There is no room for such a law in the case of normative explanation, but we need something that mediates between reasons and my doing X in the case of normative explanation too: It is a normative judgment that does. As we saw, the presence of laws is an essential feature of nomological explanations, and so normative explanations must be a kind of explanation fundamentally different from nomological explanations.

Now there is a straightforward sense in which normative explanation understood in this way is essentially subjective: It yields an adequate explanation only when the agent applies it to his own action. Suppose that I apply (B) to understand your action. That is,

(F) You were in circumstances of kind C.

The appropriate thing to do in C was X.

Therefore, you did X.

But this will fail to be an adequate explanation for your action (unless other substantial premises are supplied). There are two reasons for this (both noted by Kim), corresponding to the two premises of the explanation. The first premise of (F) expresses what I (that is, the explainer) believe. But it cannot explain your action unless you yourself believe it. You yourself may not believe that you were in circumstances of kind

110 As Kim says, “what does the explanatory work is the fact that in the circumstances the thing done was the appropriate thing to do” (2010a, p. 136).
C. But there is a more fundamental reason why normative explanation is essentially subjective, which relates to the second premise. As we saw, certain considerations will not justify an agent’s doing X, unless the agent himself judges those considerations to be reasons for doing X. And those considerations do not explain his doing X either, unless he judges them to be reasons for doing X. But if we look at (F), the normative judgment is not the agent’s, but the explainer’s. It is perfectly possible that you yourself judge that the appropriate thing to do in C is something different.111

We now have a plausible characterization of what Nagel calls “intentional explanation,” or normative explanation in our terminology. Nagel claimed that intentional explanation is “comprehensible only through my point of view.” And that claim is also vindicated; the subjectivity of normative or intentional explanation is a direct consequence of Kim’s principle and normativity involved in practical reasoning. And we also see that normative explanation is fundamentally different from nomological explanation, in that it doesn’t involve any empirical law. Before moving on, let me make two more points about normative explanation.

Kim seems to think that his model of normative explanation yields a general model of action explanation. As we saw, normative explanation is essentially subjective, but we can apply it sympathetically. He says, “a third-party's understanding of an agent's action is to be able to recapitulate, or re-enact, the agent's deliberation leading to the action.”

111 There is one potentially important point that Kim neglects. He seems to think that one can always explain one’s own actions by normative explanation. But one’s beliefs and standard of appropriateness may well change over time. For example, I once believed that it would rain, but now I don’t. Or I no more judge raining to be a reason to take an umbrella. Then explaining my action will be just like explaining other people’s actions.
Kim sometimes talks as if nomological explanation is never an adequate kind of explanation for human actions. But I think that this can’t be right. It seems to me undeniable that we often explain people’s actions (including my past actions) by appealing to some general truths about how the agent involved usually behaves in such and such situations. Moreover, human actions are a part of the physical world, and if nomological explanation is an adequate kind of explanation at all, it should apply to actions as well. The point, I think, rather should be that there are two quite different kinds of explanation that are appropriate for actions.

Another point worth mentioning is the relationship between explanation and “prediction.” It is well known that in the case of nomological explanation, there is a symmetry between explanation and prediction. That is, the argument that constitutes an explanation can be employed to predict future events. This is a natural consequence of the fact that in nomological explanation, the explanans afford “good grounds to believe or expect” that the explanandum happened. Now I think that something similar holds in the case of normative explanation. How do I know about my own future actions? For example, how do I know whether I will take an umbrella with me? In order for other people to know it, they may have to appeal to some general information about how agents act in certain situations. However, as many philosophers have observed, it would be very much odd if I went through a similar procedure. I can know it simply by making a decision, or by going through deliberation process. There is a symmetry between the process whereby I explain my actions and the process whereby I know about my future action. It would be misleading to call such knowledge a “prediction,” but there seems to
be a certain parallel between nomological explanation and normative explanation in this regard.

4. The Internal Perspective and Normative Explanation

The idea that we take two different perspectives or points of view on action is intuitively attractive, but it is something that needs to be explicated. As we saw, Nagel seems to think that the internal point of view on action is somehow essentially connected to what he calls “intentional explanation,” or normative explanation. But exactly how are they related?

Nagel was not very explicit about this. I want to make the following proposal: The practice of normative explanation itself is constitutive of the internal point of view. That is, taking the internal point of view on my own action consists, at least partly, in taking it to be explained by normative explanation. I won’t claim that it is all there is to the internal point of view; perhaps there are some other things involved in the internal point of view on action. But I think that it is very much central to it. We often say that from the internal point of view, the agent explains his own action by referring to his justifying reasons. But if I am right, “from the internal point of view” is in fact redundant; for to take one’s actions to be a matter to be explainable by justifying reasons is to take the internal point of view toward those actions.

Initially, it seemed very mysterious what those internal and external perspectives amount to. But there is nothing mysterious about the fact that there are certain explanatory practices we employ to explain our actions (although Nagel thinks that there
is something mysterious about such explanation). So if I am right, then we come to have a reasonable explication of what the internal perspective consists in.

I think that it is intuitively not implausible to think that those explanatory practices constitute perspectives in the relevant sense. Recall our analogy of two people seeing a table from different perspectives. They experience the table differently, but both of them may be representing it correctly, from different perspectives. Now we can say something similar about the two perspectives on action. When I consider my action from the inside, I explain and understand it by referring to justifying reasons. But as you consider my action from the outside, you explain and understand it by using your nomological information. The two ways of explaining actions are different, but they may be equally legitimate ways of explaining and understanding my action.

However, the last remark needs defense. There are two concerns. Nagel seems to doubt that a normative explanation is really an adequate explanation of action. Another concern is the so-called problem of “explanatory exclusion.” Can there really be two different but equally adequate explanations of one phenomenon? I will discuss Nagel’s concern in the next section. After that, I will turn to the second concern.

5. Intelligibility of Normative Explanation

Nagel believes that he is led to conclude that “it is impossible to give a coherent account of the internal view of action.”¹¹² It is because he thinks that normative explanations of actions are “either incomprehensible or clearly inadequate.” But is he right?

¹¹² Nagel (1986), pp. 112-3.
Let’s look little more closely at why Nagel thinks that it is not an explanation at all.

“[I]ntentional explanations must simply come to an end when all available reasons have been given, and nothing else can take over where they leave off. But this seems to mean that an autonomous intentional explanation cannot explain precisely what it is supposed to explain, namely why I did what I did rather than the alternative that was causally open to me. It says I did it for certain reasons, but does not explain why I didn’t decide not to do it for other reasons. It may render the action subjectively intelligible, but it does not explain why this rather than another equally possible and comparably intelligible action was done.”

Nagel’s point seems to be this. Suppose that I explain my doing X in terms of reason R. But R can explain my doing X only if R “necessitates” X. But by the nature of normative explanation, there is no such necessitation. I could do something different X* for the same reason, if I judged R to be reason for X* rather than X.

In responding to this, Kim says:

“Nagel himself admits that this account makes the choice and action “subjectively intelligible.” Subjective intelligibility, if it means intelligibility to the agent, is all we need or should want. Nagel goes on to complain that this explanation does not explain why this action, rather than another one, was chosen. But that is to complain that the agent's way of making her action intelligible to herself doesn't yield an objective, perhaps Hempelian, explanation. Well, so be it!”

I think that he is basically appealing to the fact that we do have such a practice of explaining actions. An explanation is given as an answer to “why”-question, and retracing the deliberative process is precisely how we actually answer the question, and as a result, we gain good understanding of actions. What else do we need to say?

---

114 Kim (2010a), pp. 140-1.
Kim of course is well aware that there is a deeper problem, which concerns the nature of explanation and explanatory relation. In virtue of what relation does the explanans of normative explanation (that is, justifying reasons) explain the explanandum? As we saw earlier, in the case of nomological explanation, we have reasonable (candidates of) answers to the corresponding question: *nomic expectability* and *causal relation*. It seems to be a common ground between Nagel and Kim that normative explanation fails to meet either of these conditions. Nagel thinks that this undermines intelligibility of normative explanation, but Kim denies that. Here is another relevant passage from Kim:

“As many have noticed, ‘explanation’ has a double meaning, especially as regards human behavior: sometimes it simply concerns the causal provenance of the thing to be explained, a recounting of why and how it has come to pass, but sometimes it carries the sense of ‘justification.’ ‘Explain why you did A’ uttered by the boss to a subordinate in a certain tone of voice is not a request for a causal explanation of A; it is a request for a justification or defense.”

Kim seems to be saying that explanation is a “disjunctive concept.” In some cases, explanations explain in virtue of causal relations, but in other cases, explanations explain in virtue of justification.

I want to meet Nagel’s challenge in a different way. I want to object to an assumption both Nagel and Kim apparently subscribe to; that is, the assumption that normative explanations fail to meet the requirements that nomological explanations meet; that is, *nomic expectability* and *causal relation*. I claim that we can reasonably construe normative explanation as satisfying the core ideas behind both of these requirements.

---

What is behind the requirement of “nomic expectability” is the idea that explanans should offer “good grounds for expecting or believe” that the explanandum happened. Empirical laws are supposed to be necessary and informative, and so can play such a role very well. Then what is more fundamental in the idea of “nomic expectability” may not be the involvement of laws per se, but rather the epistemic justification that it, in conjunction with statements about initial conditions, can transfer to the explanandum. But then, don’t the explanans of normative explanations also afford good epistemic grounds, at least for the agent himself, to believe that the action occurred? This point is reflected in the fact that the deliberation process can give the agent not only explanation but also knowledge about one’s future action. That is, as I said in an earlier section, we go through the process of deliberation to know about our future actions. In fact, self-knowledge of one’s own actions acquired in that way is supposed to be more secure than the third person knowledge based on nomological information. There is no “nomic” expectability in the case of normative explanations, but a similar sort of expectability seems to hold for them.

Let’s turn to the second requirement of causality, which is more important for our purpose. It is certainly true that the explanans of an normative explanation (that is, justifying reasons) can’t be seen as describing the cause of the explanandum. They in no way causally necessitate my doing X, as the explanans of nomological explanations are expected to do. Nonetheless, I want to hypothesize that we can reasonably construe normative explanations “realistically,”116 and regard them as giving information about the

116 Here, when I say I construe normative explanation “realistically,” I have in mind something like explanatory realism, which Kim discusses in his papers on explanation. Roughly,
“causal provenance” of the action. Recall that a normative explanation explains an action by “[retracing] the process of deliberation in which he engaged and which resulted in the formation of decision that led to the action.” By giving such an explanation, I am in effect saying that it’s how the action came about: I determined it by considering those reasons, through the process of deliberation. Instead of antecedent events as causes, I am invoking *myself* as the cause of the action. In other words, we can interpret normative explanation as a sort of *causal explanation* appealing to *agent causation*. Let me put the hypothesis as follows:

**Realistic Interpretation of Normative explanation:** A normative explanation of an action, just like a nomological explanation of it, explains it by tracking causal provenance of action. However, the kind of causal relation tracked by a normative explanation differs from that tracked by a nomological explanation; it is agent causation, not event causation.

I believe that there are good reasons to construe normative explanation realistically in this sense. First, let me quote Kim’s passage again:

“‘Explain why you did A’ uttered by the boss to a subordinate in a certain tone of voice is not a request for a causal explanation of A; it is a request for a justification or defense.”

---

the idea is that “explanations track dependence relations” (Kim, 1994), p. 184). What he means by dependence relations, at least primarily, seems to be relations between events. But if we interpret “dependence relations” broadly so as to include something like agent causation, my position here can be called “explanatory realism about normative explanation.”

It is true that the boss may ask for justification for his action, by asking “Why did you do A?” But the boss may ask the same question truly wondering how it really has come to pass, or how it causally originated (with the same words, “Explain why you did A,” but perhaps in a different tone of voice). Knowing the boss’s intention well, what would the subordinate do? Exactly the same; that is, he would retrace his deliberation process that led him to doing A. That’s precisely how he would think that his action had come to pass. He would be trying to give information about the causal provenance of his action. He in effect would be saying that he himself determined the action by considering reasons.

A related but different point: Suppose that the subordinate actually decided to do A for reason R. But when asked for an explanation, he comes up with different reason R*, which justified doing A, but didn’t actually motivate him to do A (perhaps to mislead the boss, or even himself). Then surely, we will say that his explanation is incorrect. The correct explanation is that he did it because of R. But the fact that we can talk of “correct” or “incorrect” explanations implies that we are committed to a certain metaphysical relation that makes it correct. What could that relation be? A plausible answer is that it is the relation of agent causation. The correct explanation is correct because that’s how the subordinate caused his action.

An interesting perspective emerges from our discussion. It is often thought that our concept of (event) causation originated from our explanatory practice. Perhaps (event) causation is something that is posited as a relation that grounds the explanatory relation. If this has some plausibility, we can say exactly the same about agent causation. Once we

119 Cf. Kim says: “a realist conception of explanation holds that such notions as “objective truth” or “correctness” or “accuracy” make sense for explanations, and do so in a more or less literal way” (Kim 1989, p. 94).
interpret normative explanation as a realistic explanation, it might be equally plausible that the concept of agent causation originated from our practice of normative explanation. That is, it is the kind of causation that grounds normative explanation. Would this amount to saying that agent causation is unreal? Not necessarily. Just as the fact that the concept of event causation is originated thus doesn’t imply that event causation is unreal, I think, so the fact that the concept of agent causation is originated from such practices of normative explanation does not imply that agent causation is unreal.

This, I think, is something that can be accepted even by those (perhaps including Nagel and Kim) who think that it is a mistake to construe normative explanation realistically. However, just because they interpret it unrealistically, they might think that the concept of agent causation is ill-founded and incoherent. But as I have just claimed, I don’t see any good reason not to interpret it thus. In any case, we now see that there is a sense in which the concept of agent causation is essentially subjective.

This completes our discussion of the three groups of questions, which I referred to as (I), (II), and (III) in Section 2. In the next section, I will consider another problem that naturally arises from our discussion so far.

6. A Worry about “Explanatory Exclusion”

Earlier in this chapter, I expressed the hope that by viewing the two conceptions of actions as a matter of different perspectives, we may be able to reconcile them. If I am right that two perspectives on actions are a matter of different explanatory practices, then don’t we have a satisfactory reconciliation of the two?
Unfortunately, the issue of incompatibility reemerges at the level of explanation. Is it really possible that both a nomological explanation and a normative explanation are adequate explanations of an action? This is an instance of the general problem of explanatory exclusion. Take a simple example for illustration. I hit Jones. Why? Suppose that two explanations are given: “Because he insulted me,” and “Because I hated him.” Can we really accept both as explanations of my hitting him? No, unless the explanans of the two explanations are somehow related (perhaps I hated him because he insulted me), or they only jointly explain my hitting him (it is because he insulted me and I hated him). So Kim states the principle of explanatory exclusion as follows: “No event can be given more than one complete and independent explanation.”

Likewise, then, when I explain my action by referring to my justifying reasons, and you explain it by referring to its causes and psychological laws, can both of us be really right? As Kim points out, “the incompatibility between these explanations stands out in the starkest way when they are both construed as causal explanations.” So for those who construe normative explanation as non-causal (as Kim and Nagel do), this may not be a serious problem. (But they face different problems, as I argued in the previous

---

120 See Kim (1988) and (1989). This problem was originally introduced in the context of the debate of compatibility of “mechanistic explanation” and “purposive explanation” of actions. As I understand the debate, “purposive explanation” here is understood as nomological or causal explanation, and so this particular debate is not directly relevant to our issue here.


123 Kim himself suggested that normative explanation might avoid the exclusion problem. (See Kim (1989), p. 103, note 4.) But he also says that merely taking a non-realist view about explanation doesn’t completely avoid the exclusion problem. He says: “Even if we abandon the idea that there are objective explanatory relations in the world, we may still find something cognitively unsettling and dissonant about having to face, or accept, two or more independent explanations of the same phenomenon” (p. 96).
section.) But to those who regard it as a kind of causal explanation, the problem of explanatory exclusion does arise “in the starkest way.”

How should we respond to this concern? There are two obvious routes, corresponding to the two provisos of the principle of explanatory exclusion. One is to deny completeness of one or both kinds of explanation. Perhaps a nomological explanation of my action only partly explains my action, and it has to be complemented by normative explanation to be complete. Or putting the point in terms of causation, my action is not wholly determined by event causation, but partly by myself. If one takes this position, it will be unavoidable to view agent causation as a *sui generis* kind of causation. As I have emphasized all along, however, I want to avoid such a view, if possible.

Another obvious option is to deny the independence of the two kinds of explanation. This position can be motivated by the fact that the situation is parallel to other instances of the exclusion problem often discussed in the literature: for example, biological explanation and physical explanation of a certain biological phenomenon. They are not mutually exclusionary, because there is a sense in which biological explanation is dependent on physical explanation; the former is reduced to the latter. Then why not do the same as for nomological explanation and normative explanation?

My answer to the explanatory exclusion problem then is that it can be avoided by reducing agent causation to event causation. So we come back to the reductionist project that we left unfinished in the last chapter.

124 Randolph Clarke can be seen as taking such a route. He locates agent causation in the context of normative explanation in our sense, and seems to think that causal explanations of an action by mental states are incomplete. See Clarke (1993), esp. pp. 290-1.
7. Conclusion

Let me summarize our discussion in this chapter. My primary purpose was to give an account of what the internal perspective on action consists in. For that, I first considered what I called normative explanations of actions, which explain them by referring to justifying reason, and are essentially normative and subjective. And second, I claimed that the internal perspective is constituted by such an explanatory practice. Third, I claimed, in opposition to Nagel, that we can make good sense of normative explanation, by construing it realistically or causally. These points together, I think, give a fairly coherent account of what the internal perspective consists in. In the last section, we considered the problem of explanatory exclusion that arises when we construe normative explanation realistically, and argued that this problem can be resolved by reducing agent causation to event causation.

So we naturally came back to our reductionist project. Moreover, I think that we are now in a better position to address some of the questions we raised in the last chapter. I turn to them in the next chapter.
CHAPTER 5

Conclusion: Reduction of Agent Causation

1. Introduction

In Chapter III, I favorably considered the possibility of applying the functional reduction model to agent causation. I tried to identify potential problems with this project. One problem was that we needed to make sense of distinct “levels” to which agent causation and event causation belong, in order to apply the functional reduction model to agent causation. Another was to identify the role of agents in producing actions, which is necessary to get the process of functional reduction going.

In the last chapter, I considered what the internal perspective on actions consists in, and claimed that it is constituted by our way of explaining our own actions by referring to justifying reasons—that is, normative explanation. I claimed that the concept of agent causation is closely tied to such explanatory practice.

In this final chapter, I want to pull together the discussions in the last two chapters, and complete our project. First, I will claim that the existence of two distinct explanatory practices is enough to generate “levels” that are required for the project of functional reduction. Second, I want to claim that the role of an agent in producing actions needs to be found in the context of normative explanation. What naturally emerges from this is that an agent’s role in producing an action is to make a normative judgment about what is
an appropriate thing to do in a given situation. And I want to make some suggestion about what could be mental states or events that realize this role. My suggestion will remain sketchy, but if it turns out to be plausible, we will have come a step closer to the desired reconciliation of agent causation with event causation. I will conclude the chapter with some closing remarks.

2. “Levels” as Explanatory Levels

I think that in order to reconcile agent causation with our naturalistic view of the world, it is very important to see that our conception of agent causation belongs to a different “conceptual scheme,” “level of descriptions” or “domain,” than our conception of event causation. First, that may be what makes agent causation conceptually unanalyzable into event causation. Second, the idea of ontological or functional reduction seems to require what is reduced to be at a different level from what does the reducing. But precisely what does it mean in general to say that there are different levels or domains to which certain phenomena belong? Does it really make sense that agent causation and event causation belong to different levels or domains?

Unfortunately, we don’t seem to have a well-articulated view about what this talk of levels amounts to.\(^\text{125}\) In some cases, it seems to be assumed that what constitute different levels are different kinds of things. For example, elementary particles constitute the physical level, molecules the chemical level, and cells the biological level, etc. Sometimes, what is more fundamental seems to be different sets of properties. For

\(^{125}\) For a sustained discussion of this issue, see Kim (2002).
example, psychological properties, such as feeling pain or believing something belong to the psychological level, while properties such as undergoing C-fiber firing belong to the physiological level. Or one might think that it is a hierarchy of different theories that constitute levels. Corresponding to the psychological level, there is the psychological theory, and corresponding to the physiological level, there is the physiological theory.

However, all these don’t seem to directly apply to the case at hand. We say that an agent’s action is caused by his beliefs and desires, and we say that an agent herself brings about an action. Are there distinct kinds of things or properties involved in the latter? It is hard to think so. In both cases, we are interested in an agent, and her actions. And it may not sound right that different theories are involved either. When we say that an agent’s action is caused by beliefs and desires, we may be invoking a certain psychological theory. But what about agent causation? Is there a separate psychological theory that explains action by agent causation? One may well doubt that there is such a theory.

For all this, I think that there is a straightforward sense in which agent causation and event causation belong to different levels. There are two distinguishable explanatory practices that are tied to different kinds of causation. We sometimes explain people’s actions by subsuming them under psychological laws; that is, we explain them nomologically or causally, by invoking event causation. Other times, especially when we explain our own actions, we employ a completely different kind of explanation. We explain them by justifying reasons, and if what I claimed in the last chapter is right, this sort of explanation explains actions by invoking agent causation. I think that those
different kinds of explanation can be regarded as generating distinct *levels* that are required for the reductionist project to get going.\(^{126}\)

I don’t think that there is anything arbitrary with this. On the contrary, I think that it is not implausible to think that in general, what generate different levels are our explanatory practices; that is, levels in the relevant sense are “explanatory levels.” Corresponding to the psychological level, there are psychological explanations of certain phenomena. They explain those phenomena by referring to certain psychological *properties* and *things*. Moreover, psychological explanations will explain those phenomena by appealing to certain psychological *theories*. And *mutatis mutandis* for other levels, such as the physical level or the biological level. So perhaps in general, what is fundamental to the level talk is not different sets of properties or things, or different theories *per se*, but different explanatory practices.

I admit that all this is very rough, and needs to be worked out further. But I think that it is enough to get the process of ontological reduction of agent causation going. Let me now turn to how all this helps completing the reductionist project.

### 3. The Role of Agent and Its Realizer

As we saw in Chapter III, the functional reduction of agent causation is supposed to go in the following two steps:

1. Agent S plays role F in producing action A.

\(^{126}\) As Kim emphasizes, any talk of “levels” involves the idea that one level is “higher” or “lower” than another (Kim 2002, p. 43). But in what sense could the level of agent causation be “higher” than the level of event causation? Well, if I am right about the reductionist project, then there is a straightforward sense: The former is *reducible* to the latter.
2. Mental state M plays role F in producing action A.

3. Therefore, S’s causing action A = M’s causing action A.

If we complete step 1 and step 2, then reduction of agent causation to event causation is derived as a logical consequence. The crucial task then must be to identify the role of agents. As we got clearer on the “level” to which agent causation is supposed to belong, I think, we are in a better position to identify the role of agents.

I claim that the role of agents should be found in the context of normative explanation. Agent causation is something that is visible only from the internal perspective. And viewing an action from the internal perspective is a matter of taking it to be explainable by normative explanation. If this is right, then the only way we can identify the role of an agent in producing actions must be by looking at what role normative explanations of actions attribute to the agent.

Let me formulate normative explanations in the following way:

\[ R \]

Therefore, I did X.

Here R is supposed to be the agent’s justifying reasons for doing X. For example, I accepted the job offer A, because it pays more than my present job. Or I took an umbrella because it was going to rain and I didn’t want to get wet. As we saw, this sort of explanation explains actions by “retracing” the agents’ deliberation processes leading to the actions. That it was going to rain, and that I didn’t want to get wet explain my action, because I decided to take an umbrella by considering those reasons.

What is the role of an agent in this picture? Recall Nagel’s complaint about the adequacy of normative explanation. It was that justifying reason R fails to necessitate my
doing X, and hence that this explanation doesn’t really explain why I did X. Nagel is quite right that there is a certain “gap” between justifying reason R and my doing X, as the former fails to necessitate the latter. But this is exactly what points to a solution to our problem. I want to claim that we can identify the role of the agent as *whatever fills in this gap*. In other words, the role of the agent must be to *mediate between those justifying reasons and actions* to make normative explanation genuinely explanatory.

What could this be more specifically? Recall that normative explanations should always be mediated by *normative judgments* to the effect that in a given situation, doing X is an appropriate thing to do. Justifying reason R can adequately explain an agent’s doing X, only if the agent *judges* R to be reason for doing X. If I judged R to be a reason for doing something different, then it would cease to justify and explain my doing X. So I want to claim that seen in the context of normative explanations, the role of an agent must be *to make normative judgments about what is an appropriate thing to do in a given situation by reflecting on reasons*.

I should emphasize that all this is a very natural consequence of our discussion so far. Agent causation is something visible only from the internal perspective, and the internal perspective on actions is constituted by the practice of normative explanations. So in order to find the role of the agent, we have to look at normative explanations of actions. And in that context, the role that the agent plays should be nothing but to make a normative judgment about what is an appropriate thing to do in a given situation.

Let’s turn to step 2 of the functional reduction. What we have to do in step 2 of reduction is to *find the mechanism that plays the role of the agent*. Then assuming that the role of an agent is to make a normative judgment about what is an appropriate thing...
to do in a given situation, we should find the mechanism that produces such a normative judgment. Could there be such a mechanism at the “base level”? That is, could there be mental states or events that play the role of the agent thus identified? One might insist on a purely a priori ground that it is the agent himself, not any mental states or events, that could produce such judgments. But I don’t see any ground for such a claim. Or it should be at least an empirical question (perhaps of empirical or commonsense psychology) whether there is such a mechanism (constituted by mental states and events) that produces such a normative judgment.

So specifically what could be such mechanisms that play the role of an agent? In other words, what is the realizer of the agent’s role? I can give only a rough and speculative suggestion. An agent reaches a normative judgment about what is an appropriate thing to do in a given situation through practical deliberation processes. So I think that what plays the agent’s role must be the mechanism that drives practical reasoning.

It is interesting that David Velleman reaches a similar conclusion, through a route that is somewhat different from mine. As we saw, he tried to identify an agent’s role by looking at the role that “ordinary parlance attributes to him.”¹²⁷ He writes:

“We say that the agent turns his thoughts to the various motives that give him reason to act; but in fact, the agent’s thoughts are turned in this direction by the desire to act in accordance with reasons. We say that the agent calculates the relative strengths of the reasons before him; but in fact, these calculations are driven by his desire to act in accordance with reasons. We say that the agent throws his weight behind the motives that provide the strongest reasons; but what is thrown behind those motives, in

fact, is the additional motivating force of the desire to act in accordance with reasons.”\textsuperscript{128}

In short, as Velleman sees it, the agent’s role in producing actions is to reflect on various considerations and “[to] calculate relative strengths” of them, and determine to act on particular reasons. And he concludes that what realizes this role is “a desire to act in accordance with reasons.” Here is a little more detailed picture:

“What mental event or state might play this role of always directing and never merely undergoing such scrutiny? It can only be a motive that drives practical thought itself. That is, there must be a motive that drives the agent’s critical reflection on, and endorsement or rejection of, the potential determinants of his behavior, always doing so from a position of independence from the objects of review. Only such a motive would occupy the agent’s functional role, and only its contribution to his behavior would constitute his own contribution.”\textsuperscript{129}

I think that there is something right about this proposal. As I said, what realizes the agent’s role is what drives practical reasoning. And an important part of it may be a motive “that drives practical thought itself.”

As I said, all this is very sketchy. But in the functional reduction model, this is all we can do as philosophers. For, in general, finding the realizer of the role of what is reduced (that is, step 2 of functional reduction) is supposed to be an \textit{a posteriori} and empirical work.

\begin{flushright}
\textsuperscript{128} Velleman (1992), p.141.  \\
\textsuperscript{129} Velleman (1992), p.139.
\end{flushright}
4. Wrapping Things Up

We started by taking our intuition and phenomenology that we have about our own actions at face value. The core idea of agency lies in the fact that we ourselves bring about, or cause, our own actions. As far as we are committed to this idea, most of us seem to be committed to agent causation.

But agency and agent causation are problematic, because it apparently conflicts with our naturalistic view of the world. According to this view, every event, if it is caused at all, is completely caused by antecedent events. The naturalistic worldview doesn’t seem to allow anything other than event causation. So the problem that agency raises is to resolve this apparent tension, or to reconcile agent causation with event causation.

One obvious way to achieve this is to reduce agent causation to event causation. So I have examined the possibility of such a reduction. We considered two different proposals for reducing agent causation to event causation. The first is to conceptually reduce agent causation to event causation. The idea is that we can paraphrase statements about agent causation into statements about event causation. Some philosophers seem to have thought that this would be the way that reduction of agent causation had to go, if reduction of agent causation to event causation were possible at all. However, there is a virtual consensus that such conceptual reduction is not possible. I tried to track down reasons for this, and suggested that the main reason for conceptual irreducibility is that agent causation and event causation belong to different “conceptual schemes.” There is something essentially subjective about the concept of agent causation, while event causation is completely objective.
But the very fact that they belong to distinct conceptual schemes seemed to open the possibility of another kind of reduction. For it makes the case of agent causation and event causation parallel to other familiar cases of *ontological* reduction. For example, we don’t expect statements about pain to be translatable into statements about C-fiber firing, as they belong to completely different conceptual schemes, but nonetheless, it is not unreasonable to hope that the former is ontologically reducible to the latter. I considered the possibility of applying the most popular model of ontological reduction, the “functional reduction model,” to agent causation. Despite some apparent obstacles, I argued that this model of reduction could be fruitfully applied to agent causation.

What emerged from our discussion of the two modes of reduction was the idea that agent causation and event causation belong to different “conceptual schemes” or “levels.” I claimed that what constitute these distinct schemes or levels are two different *perspectives*—internal or subjective, and external and objective—on actions. And the latter in turn are constituted by two different explanatory practices. We may explain our actions by subsuming them under causal laws, possibly involving mental states and events. But that’s not how we normally explain our own actions. We explain our actions by referring to justifying reasons. And I argued that the concept of agent causation is closely tied to such an explanatory practice.

What is crucial for ontological or functional reduction of agent causation to event causation is to identify the role of an agent in producing actions. I argued that once it is made clear that agent causation is visible only from the internal point of view, or in the context of normative explanation, it is relatively easy to find the role of an agent in such explanations. The agent’s role is to make a normative judgment about what is an
appropriate thing to do in a given situation. And I claimed that what plays this role in the base domain must be the mechanism that contributes to the agent’s normative judgment about what is an appropriate thing to do in a given situation. If this is right, then in order to understand agency and agent causation, what we need to understand is the mechanism of practical reasoning.

Before closing, let me quote one of the passages I quoted earlier with respect to the problem of “disappearing agency” once more:

“All conceiving of action as the result of a chain of events that are causally connected removes the agent from the picture altogether. Being the conduit for a causal chain is a passive affair; being an agent is being active. The former cannot possibly exhaust what is involved in the latter.”

What can we say about this sort of concern now? An agent’s activity is something that can be seen only from the internal point of view, and it consists in the agent’s playing a certain role in explaining actions from that point of view. But what realizes that role in the naturalistic world is nothing other than mental states and events.

The objector might not be persuaded of this. He might insist that reducing agent’s activity to mental states and events amounts to “removing the agent from the picture altogether.” Such a concern may well be perfectly understandable. But I want to respond that if that is a concern, it is a quite general one that applies to any reductionist project. I think that there is a persistent feeling that reducing something amounts to eliminating it. For example, one might feel that if pain is reducible to a certain physiological state, then there is really no pain as we believe it to be. But, as Kim nicely points out, “there also is

---

130 Enç (2003), p. 3.
the apparently opposed view: to be reduced is to be legitimized.” Under that conception of reduction, our reductionist project may have contributed to legitimizing agent causation.

131 Kim, (1984a), pp. 94-5.
BIBLIOGRAPHY


