THINKING ABOUT ACTION

by

Joshua Steven Hancox

AB, University of Chicago, 2008

Submitted to the Graduate Faculty of

The Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

THE DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Joshua Steven Hancox

It was defended on

March 10, 2015

and approved by

Peter Machamer, Professor, Department of History and Philosophy of Science, University of Pittsburgh

John McDowell, Distinguished University Professor, Department of Philosophy, University of Pittsburgh

Sarah Paul, Assistant Professor, Department of Philosophy, University of Wisconsin – Madison

Kieran Setiya, Professor, Department of Linguistics and Philosophy, Massachusetts Institute of Technology

Dissertation Director: Michael Thompson, Professor, Department of Philosophy, University of Pittsburgh

THINKING ABOUT ACTION

Joshua Steven Hancox, Ph.D.

University of Pittsburgh, 2015

The promise of action theory (the study of intentional action) is that it might provide a new way into old disputes about the foundations of ethics, or the mind-body problem, or even first-order moral questions. The difficulty is accounting for the three quite different characteristics of intentional action: the characteristic way practical thought affects the world, the distinctive patterns and norms of means-end reasoning, and a special way of knowing about one's own actions. I explore the idea that we must first understand how agents think about action in order to understand these central features. In particular, I argue that practical thought – intentions and means-end beliefs – represents itself as the cause of its object. In addition to resolving action-theoretic debates, this account fulfills some of the promise of action theory, providing a rigorous foundation for a number of ethical and metaethical positions.

TABLE OF CONTENTS

LIST OF FIGURES

I would like to thank Robert Batterman, John Broome, Michael Caie, Charles Goldhaber, Anja Jauernig, Bihui Li, Suzanne Love, Peter Machamer, Adam Marushak, John McDowell, Sarah Paul, Raja Rosenhagen, Kieran Setiya, James Shaw, Elizabeth Silver, Robert Steel, Michael Thompson, the members of the University of Pittsburgh dissertation seminar, and anonymous reviewers at *Mind* and *Noûs* for their insightful and invaluable feedback on the ideas presented here. You have to write your dissertation yourself, but you don't have to write it alone.

1.0: INTRODUCTION

1.1: PHILOSOPHY OF ACTION IN THE TWENTIETH CENTURY

> "A man is pumping water into the cistern which supplies the drinking water of a house. Someone has found a way of systematically contaminating the source with a deadly cumulative poison whose effects are unnoticeable until they can no longer be cured. The house is regularly inhabited by a small group of party chiefs, with their immediate families, who are in control of a great state; they are engaged in exterminating the Jews and perhaps plan a world war. --The man who contaminated the source has calculated that if these people are destroyed some good men will get into power who will govern well, or even institute the Kingdom of Heaven on earth and secure a good life for all the people; and he has revealed this calculation, together with the fact about the poison, to the man who is pumping. The death of the inhabitants of the house will, of course, have all sorts of other effects; e.g., that a number of people unknown to these men will receive legacies, about which they know nothing.
>
> "This man's arm is going up and down, up and down. Certain muscles, with Latin names which doctors know, are contracting and relaxing. Certain substances are getting generated in some nerve fibres -- substances whose generation in the course of voluntary movement interests physiologists. The moving arm is casting a shadow on a rockery where at one place and from one position it produces a curious effect as if a face were looking out of the rockery. Further, the pump makes a series of clicking noises, which are in fact beating out a noticeable rhythm.
>
> "Now we ask: What is this man doing?"[1]

As Anscombe's story suggests, the topic of "philosophy of action" is a perfectly ordinary one: an interest in action pervades our ordinary lives. But then, haircuts are like this too. Philosophers have been drawn to the topic of *action* (but not haircuts) because of its close connections with the mind-body problem, with the content of morality, and if some more recent theories are correct, with the foundations of morality. Like perception, intentional action appears to be a place where thought makes contact with the world. We want an account that renders this somehow intelligible. The content of many moral theories, and much everyday moral theorizing, relies on distinctions between intentions and side effects, between "on purpose" and "by accident," between action and inaction. And, if one suspects that intentional action defines the topic of moral theory, then one might suspect that it also provides its foundations: and so various sorts of "constitutivist" theories have it. My aim, in this dissertation, is not to address these concerns directly, but rather to try to provide an account of their common theme: intentional action.

[1] Anscombe 1957 §23.

Philosophers have typically approached the idea of intentional action via one of four possible routes. Some have focused on the relations between the agent's thought and the world, e.g. the problem of deviant causal chains. Some have focused on the internal, i.e. instrumental or means-end, relations between intentional actions. Some have focused on the relation between action and belief, i.e. practical knowledge. And others have focused on the connection between action and "rational autonomy," the characteristically human exercise of free reason. I'll elaborate on each of these in turn.

Donald Davidson gives the classic statement of the first topic in his "Actions, Reasons, and Causes." He points out that when a person had many reasons for an action, we can ask: but *which* reason did he act on?[2] Davidson answers: the reason he acted on is the reason which *caused* his action. But this leaves us with the problem of "deviant" or "wayward" causation, exemplified by Davidson's Climber:

> "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."[3]

The problem of course is that while the climber has reasons which cause an action, they do not do so "in the right way." And while Davidson himself despaired of giving a non-circular account of "in the right way," many other philosophers have attempted solutions. Consider Miles Brand's solution: he notes that examples of deviant causal chains involve a gap between attitude and action into which deviance can creep. Thus Brand suggests that, in the case of basic actions, actions we do not perform by means of other actions, intentions cause action *immediately*: with no gap, there's no place for deviance.[4] Non-basic actions are non-deviant just in case they are caused by basic actions according to the agent's plan. I mention Brand's answer not because it is generally accepted (no answer is), but to illustrate the basic strategy of many authors: find some property of causal connections in general, which when applied to a connection between thought and the world guarantees that it is "the right way" of connecting them.

I said that the second topic concerned the internal relations between intentional actions, that is: instrumental reasoning. What is interesting about instrumental reasoning, or correlated instrumental norms, is that it seems distinctly unlike other kinds of reasoning. Anthony Kenny points out that instrumental reasoning is in some sense

---

[2] Davidson 1963 p. 9.
[3] Davidson 1973 p. 79. In fact the reference to choice is unnecessary; the same problem can of course be raised concerning the relation between choice or intention and action. And of course the problem was noted earlier than Davidson; Anscombe, for instance, prefigures this precise example when she remarks "If someone says 'Tremble' and I tremble I am not *obeying* him - even if I tremble because he said it in a terrible voice," (Anscombe 1957 §20).
[4] Brand 1984 p. 20.

the inverse of theoretical reasoning.[5] Affirming the consequent is a logical fallacy, yet the equivalent instrumental inference - moving from an end to the means that is sufficient for it - is both valid and pervasive. Conversely, while believing the necessary consequences of one's beliefs is, again, valid and pervasive, there is no practical principle which requires us to aim at the known side effects of our actions. But these instrumental norms are also not simply derivable from desire: while it is perfectly normal to desire many incompatible things, to not desire the necessary means to desired ends, and so forth, intentional action requires somewhat more consistency. The fact that instrumental reasoning and instrumental norms appear sui generis has prompted philosophers as different as Michael Thompson and Michael Bratman to postulate that intention is metaphysically distinct (Bratman holds that it is a distinct practical attitude, while Thompson seems to hold that it is a distinctly practical process), with a unique nature generating unique norms.[6] If we hold that intention, like belief, is a propositional attitude, then we cannot derive these norms from the particular content of intentions (since both intentions and beliefs may have the same content); instead, they must derive from the peculiar form of intentions: in the words of many theorists, intentions have the opposite "direction of fit" of beliefs.

The third topic concerns "the knowledge a man has of his intentional actions," what I will call practical knowledge.[7] Here, the challenge is to explain why there is a certain connection between action and knowledge or belief. For instance, typically, if I am doing A intentionally, then I know "I am doing A." And, typically, I don't know this via the usual perceptual means: unlike other people, I don't need to see myself typing to know that I am typing. One might also come at this topic from another angle. As Anscombe originally emphasized, the claim "I intend to go the opera, but I won't go to the opera" has the air of Moore's Paradox about it: both conjuncts might be *true*, but there is something problematic about asserting both together.[8] But both of these facts are restricted to intentional action. If I am dying of cancer, there's no particular tendency for me to know that I am dying of cancer. And if I hope, or want, or imagine that I will eat a curry for lunch, I may still with perfect consistency predict that I will not. So there must be something peculiar about intentional action which underwrites this close connection with belief. Perhaps the most popular way of accounting for these facts is to assert that intentions are *constituted* by the relevant beliefs; Harman, Velleman, Setiya, and others take this approach.

---

[5] Kenny 1966.
[6] Thompson 2008 p. 92; Bratman 1987 p. 10.
[7] The phrase is Anscombe's (1957 §28).
[8] Anscombe 1957 §52.

The fourth topic, rational autonomy, is perhaps somewhat less clear. Perhaps intentional action essentially involves responding to the reasons that there are (hereafter Reasons), or perhaps the ability to have beliefs about Reasons, or perhaps something else. I won't elaborate in detail on these possibilities, since it is my ultimate view that they are not essential features of intentional action. But showing this will take some doing; it must wait on an account of the previous three features. So I will only return to this topic in Chapter Six (§6.1), where I will argue that because rational autonomy is not essentially connected to the other three features of intentional action, it is not an essential feature of intentional action.

So we have three questions about intentional action:

1: What is the right relation between intention and action?
2: What explains instrumental relations between actions?
3: What is the relation between action and belief?

It is the aim of this dissertation to answer these questions. More strongly, it is the aim of this dissertation to give a *unified* set of answers to these questions. Intentional action is not a chimera stitched together from whatever parts philosophers had lying around. An account of action should *explain* why the three apparently disparate features I discussed are all aspects of one phenomenon, intentional action.[9] Much of what I will say is not terribly new. In large part, the pieces necessary to answer these questions are already in place; they just need some rearranging and reappraisal.

My view, in a slogan, is that **practical thought represents itself as the ground of its object**. By "represents itself" I mean to extend David Lewis' account of *de se* content to the idea that a representation may be centered not merely on the agent but on the representation itself. That is, practical thought attributes to itself the property of being a ground. I say "ground" to prescind from debates about whether the relation between thought and action is causal or non-causal dependence; on my view, the basic similarity – dependence – between these views is more important than the difference. Thus, to say that A grounds B is just to say that B depends on A. Lastly, "its object" varies depending on the kind of practical thought in question. Intentions represent themselves as the ground of action.[10] Instrumental beliefs represent themselves as the ground of intentions, conditional on those intentions being instrumentally efficient.

[9] Kieran Setiya makes this point (2012 pp. 290, 293), as does Anscombe (1957 §2).
[10] In fact, it is my view that we might replace "...represent themselves as the ground of action" with "...as the ground of worldly states and events," and then explain how this entails that they are also the ground of actions – processes structured by instrumental reason. I will return to this topic in Chapter Five (§5.5).

This account is both necessary and sufficient to account for each feature of intentional action. The characteristic form of the dependence of action on thought is not anything imposed from without, but is rather a matter of the agent's own conception of how their actions depend on their practical thought. What deviance, including "basic" deviance, deviates from is the agent's own conception. In order to bear this out, we must detour into contemporary debates over the concept of causation, and reclaim the idea that causal understanding is as much the property of ordinary agents as it is of professional scientists. Second, instrumental reasoning and rationality must be grounded in a conception of instrumental beliefs. I argue that the best available decision-theoretic accounts of such beliefs fail in cases where success in action depends on the agent's reasons for action. Instead, instrumental beliefs must be about the very situation of acting on them. And this implies that instrumental beliefs represent themselves as the conditional grounds of intentions. Third, while intentions are not beliefs, they do represent the world. So when they are formed by sound reasoning, i.e. reasoning from true instrumental beliefs, they allow us to know genuine states of affairs by being the ground of those states, rather than by perceiving them. Thus agents typically (but not always) have non-observational knowledge of their intentional actions.

Quite obviously, this view is similar to – and deeply indebted to – prior "cognitivist" theories of intention, such as those developed by Harman, Velleman, and Setiya.[11] In contrast to these theories, I do *not* assert that intentions are a kind of belief: there are important formal differences between intentions and beliefs. While beliefs and intentions are alike in a) being representational states which are b) "satisfied" or not insofar as their content matches the world, they have formally distinct contents.[12] Thus the differences between intentions and beliefs are not to be explained by e.g. having a different direction of fit, or being subject to different sui generis norms. Instead, the differences between intentions and beliefs are to be explained by their formally different contents.

Plainly enough, this account of practical thought and intentional action relies on two prior notions: first, the notion of grounding or dependence, and second, the idea that a thought can represent itself as standing in (inter alia) grounding relations. The next two sections of this chapter are dedicated to developing these ideas. I will discuss them again in more detail at relevant moments in the dissertation; therefore consider what follows to be an overview (and a heads-up) about what is to come.

---

[11] I should note that Velleman later revised his description of cognitivism in ways quite coherent with what I am about to say (2007 p. xix)

[12] I do not mention different "directions of fit," since I think there is no such concept. Its initial developers, Anscombe and Searle (Anscombe 1957 §§31-32; Searle 1983 pp. 4-8), were misled by a false analogy between speech acts and mental states. See Humberstone 1992 for a sustained critique of extant conceptions of direction of fit.

1.2: CAUSATION AND DEPENDENCE

I've said that practical thought essentially involves the notion of grounding – which I glossed as dependence, either causal or non-causal. Thus we must inquire into how agents think about grounding. I suggest that we should investigate this piecemeal, and first investigate how agents think about causation.

One might balk immediately. Why should an action theorist take an interest in the internal debates of philosophy of science? Can't we get on well enough with our ordinary notion of causation, and leave the details to the specialists? Unfortunately, I think the answer is "no." In my view, causation is somewhat like Neptune: it has long exerted an unnoticed influence on accounts of action. And we labor under this alien gravity at our peril; the danger lies in not recognizing the tacit influence of controversial accounts of causation on our theories of action. In §1.2.1, I suggest how one of these accounts – the "deductive-nomological" (DN) model of causation – has caused problems for philosophers of action.

In §1.2.2, I present what I take to be a better model of causation – the "causal model" approach, one developed by a wide array of philosophers, statisticians, and psychologists.[13] In my view, this is in fact the correct account of causation. But demonstrating that – demonstrating that it is superior to the other leading views – is a task far beyond the scope of this dissertation, let alone this brief section. Instead, I'll cleave to the more restricted view that that the causal model approach is a better model of *causal understanding*, of how everyday agents think about causation. And in §1.2.3, I show how to generalize Woodward's model of causation to a conception of dependence (or equivalently grounding) generally; I will rely on this model of dependence throughout the dissertation.

1.2.1: Action and causation

Early action theory was conducted in the shadow of logical positivism, and its very specific conception of what causality was. Hempel is typically credited with developing the DN model of causation. A simple version is this:

> *Deductive-Nomological Model (DN)*: a causal explanation is a logically sound proof of some empirical fact with premises including at least one law of nature.[14]

[13] e.g. Pearl 2000; Spirtes, Glymour, and Scheines 2001, Woodward 2003; Gopnik and Wellman 2012; Hitchcock 2001 and 2013; this is only a very partial list.
[14] I've condensed this definition from Hempel and Oppenheim 1948 pp. 137-138.

The two features I wish to emphasize are *DN*'s reliance on the notions of *entailment* or *necessity* and *law of nature*, all of which figure in one way or another in the following action-theoretic remarks. Here is Ryle in 1949, defining "he boasted from vanity" as

> "...he boasted... and his doing so satisfies the law-like proposition that whenever he finds a chance of securing the admiration and envy of others, he does whatever he thinks will produce this admiration and envy."[15]

And Anscombe's brief against causalism in 1957 only makes sense against this same background:

> "But does this mean that people must have notions of cause and effect in order to have intentions in acting? Consider the question 'Why are you going upstairs?' answered by 'To get my camera'. My going upstairs is not a cause from which anyone could deduce the effect that I get my camera. ... It is not that going upstairs usually produces the fetching of cameras, even if there is a camera upstairs..."[16]

And Brand in 1984 writes:

> "Causal dependency between action types is to be explained as follows: for any event types F and G, F is causally dependent on G *iff* there is a causal law that entails that it is causally necessary that, for every event x if x is of type F, then there is an event y such that y is of type G."

And indeed, we can see how the divisions between Anscombe and Davidson were shaped by this shared background assumption. Davidson, for instance, held that intentional action involved a causal connection between thought and world. Given *DN*, this means that there must be a lawlike connection between thought and world: but since so such connection is forthcoming, Davidson was driven to his anomalous monism: the theory that token mental states are identical to token physical states, despite their being no possibility of general reductive laws relating the two.[17] This, to him, seemed to be the only way of reconciling *DN* and the causal theory of action. Anscombe, by contrast, also knew that there was no lawlike connection between intention and action; this motivated her to deny that there was a (efficient) causal relationship between intention and action.[18]

Of course, no one any longer accepts *DN* in its unvarnished form. As Woodward puts it:

> "...there is less consensus on the topics of causation and explanation in philosophy than there was three or four decades ago, when the DN model was widely accepted."[19]

And yet:

> "Moreover, all this discussion has had relatively little impact on philosophers (e.g., those working in philosophy of psychology or biology) who are not themselves specialists in causation/explanation but who draw on ideas about these subjects in their own work. To the extent that there is any single dominant view

---

[15] Ryle 1949 p. 89.
[16] Anscombe 1957 §22; see also her discussion of Ryle in §13. It's worth noting that Anscombe appears to reject *DN* elsewhere (e.g. in her 1971), but at least in *Intention*, it colors many of the remarks she makes on causation.
[17] Davidson 1978 p. 87 n. 3.
[18] Her 1983 suggests this line of argument.
[19] Woodward 2003 p. 3

> among this group, it probably remains some hazy version of the DN model: it is assumed that even if Hempel didn't get all the details right, it is nonetheless correct that explanations and causal claims must in some way 'involve' or 'be backed' by 'general laws.' The question of what a 'law' is, or what 'backed' means, or how the DN model could possibly be correct, given the widely accepted counterexamples to it, are regarded as best left to specialists."[20]

And this hazy consensus, I suggest, is responsible for the continuing intractability of a famous problem: that of deviant causal chains, of finding the right kind of causal connection between intention and action. Here are two examples:

> *Pigs:* "A man may try to kill someone by shooting at him. Suppose the killer misses his victim by a mile, but the shot stampedes a herd of wild pigs that trample the intended victim to death."[21]

> *Climber*: "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."[22]

Pigs is an example of "antecedent" deviance: the causal chain deviates from normal in the public world. *Climber* is an example of "basic" deviance: the deviance creeps in between the climber's intention and his basic action of moving his body. It has long been thought that the former problem is relatively easy to solve: what happens does not accord with the agent's conception of *how* their actions cause the desired outcome. But few have been willing to make the analogous claim about basic deviance. And this, I suggest, is because they have tacitly accepted *DN*. On such a model, if the agent has a conception of how their intentions cause their basic actions, that conception must consist of either a) knowledge of intermediate neuromuscular events or b) knowledge of the natural laws which connect intentions and actions. But neither of these seem at all plausible. I suggest that the resolution to the problem of deviance is surprisingly simple: give up *DN* and the associated picture of causal understanding. Agents do have a causal understanding of the connection between their actions and intentions; this understanding typically consists in knowledge of the circumstances in which an intention to do A will succeed. I develop this idea in more detail in Chapter Two (§2.2), but one might initially respond: isn't this ad hoc? Isn't it best philosophical practice to simply adopt the best available account of causation, even if that makes trouble for action theory? I respond: in fact, this thesis about causal understanding is backed by a leading contemporary theory of causation: the causal modeling approach.

---

[20] Woodward 2003 p. 4
[21] Davidson 1973 p. 78. Davidson credits the example to Daniel Bennett.
[22] Davidson 1973 p. 79. In fact the reference to choice is unnecessary; the same problem can of course be raised concerning the relation between choice or intention and action.

1.2.2: Woodward's "interventionist" model of causation

James Woodward's *Making Things Happen* develops an "interventionist" interpretation of the contemporary "causal model" approach to causation. I won't attempt to defend this conception of causation against its competitors, such as Salmon's causal-mechanical model, Kitcher's unificationist model, or Lewis' counterfactual reduction. Instead, I will simply take Woodward's criticisms of these views for granted, and leave a detailed defense of the causal model approach to the specialists. The details aside, I take it that there is a basic contrast between the causal model approach and these others: they are typically motivated by overtly metaphysical concerns. The basic ambition is to ensure that causation has a secure metaphysical foundation, that it will find a place in final science or final metaphysics. But this ambition means that such accounts tend to be rather far removed from ordinary life. In this section, I will suggest that Woodward's account is far better placed to account for the kind of causal understanding possessed by ordinary human agents.

Here's a metaphysical slogan: causation is a relation between particular events grounded in natural laws. For instance, L.A. Paul and Ned Hall, in their recent *Causation: A User's Guide*, begin by "recogniz[ing] certain basic facts about causation"; first on the list is "Causation seems, at least in the first instance, to relate *events*." A moment later, they claim "...causation is law-governed: minimally, what causes what is fixed not merely by what happens, but also by what the *fundamental laws of nature* are..."[23] I follow Woodward not merely in holding that these should not be the very first facts about causation one considers, but in holding that they are not facts at all: causation involves relations among variables of many different kinds, and hardly involves a notion as demanding as *natural law*. I of course utterly lack the space to reproduce Woodward's compendious arguments against rival views; in this section, I merely aim to lay out the conception of causation that I will be taking for granted.

In particular, here are three interconnected theses about causation that I will rely on. The first is that causal knowledge is knowledge of dependences among variables of various sorts. In particular, causal knowledge does not consist solely of knowledge of sequences of events. Thus Woodward writes:

> "...any explanation that proceeds by showing how an outcome depends (where the dependence in question is not logical or conceptual) on other variables or factors counts as causal."[24]

Thus, agents need not be able to identify physical laws or intervening events in order to count as having causal understanding:

[23] Paul and Hall 2013 p. 4, emphasis mine.
[24] Woodward 2003 p. 6. See also pp. 13-14.

> "We have at least the beginnings of an explanation when we have identified factors or conditions such that manipulations or changes in those factors or conditions will produce changes in the outcome being explained."[25]

The second thesis is that knowledge of invariances is causal knowledge. "Invariance" plays the role in Woodward's system that "law of nature" plays in traditional accounts: an invariance is the general fact which relates different variables. Woodward's account of them has essentially two parts. The first is a formal condition on invariances: they have to be genuinely change-relating, i.e. capable of answering at least some what-if questions. So, for instance, the (true) claim that "All igneous rocks have mass greater than zero" is not change-relating, since it tells us nothing about what a change in one of the variables (type of rock, mass) would produce in the other.[26] The second is a substantive condition: the invariance has be true for at least some values of the variables in question; that is, it needs to be "stable" or "robust" under certain changes. Woodward writes:

> "It is crucial to the causal or explanatory status of a generalization that it would continue to hold under some interventions on the values of the variables figuring in the relationship,"[27]

Notably, these are *not* changes in background conditions: a genuine invariance can be invariant in a very small range of such conditions. Rather, what matters is whether the invariance holds when we manipulate the variables it specifically is about.[28]

Third, shallow causal knowledge is possible:

> "Consider the claim (1.5.1) 'Depressing the gas pedal on my car causes it to accelerate.' Assuming that my car is functioning normally, the manipulationist view that I advocate judges that this is a true causal claim and one to which I might appeal to explain why my car has accelerated on some particular occasion. Nonetheless, as the econometrician Haavelmo (1944) observed decades ago, there is an obvious sense in which it is explanatorily shallow in comparison with the sort of theory of the internal mechanism of the car that might be provided by an automotive engineer; I don't have any very deep understanding of why my car moves as it does, if I only know (1.5.1)."[29]

We can illustrate all three of these ideas with a familiar example: a flagpole is casting a shadow. Here, Θ represents the angle of the sun, H the height of the

[25] Woodward 2003 p. 10
[26] Woodward 2003 p. 246.
[27] Woodward 2003 p. 249.
[28] Woodward 2003 pp. 248-249.
[29] Woodward 2003 p. 18

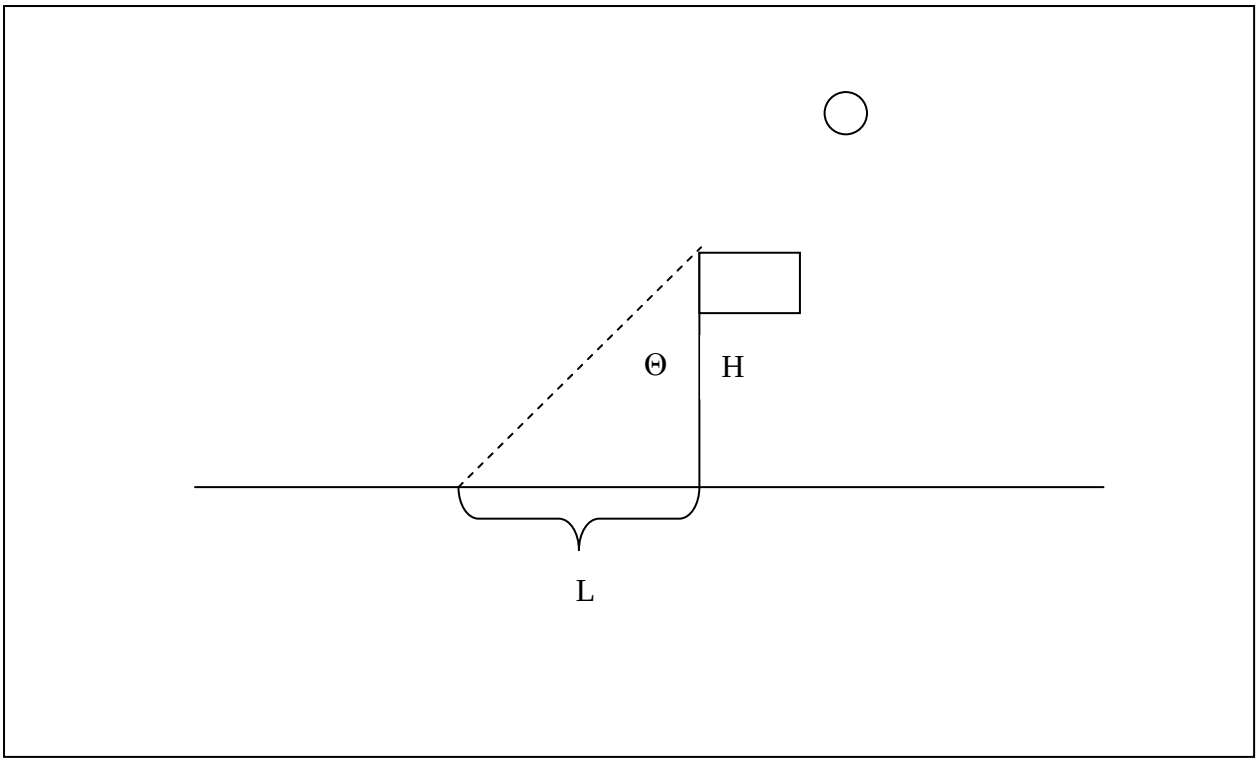

Figure 1

flagpole, and L the length of the shadow cast by the flagpole. We can express the relationship among these values as

$$L=H\tan\Theta$$

As Woodward emphasizes, this equation does not adequately represent the causal structure of the example. This is because while we can derive the height of the pole from the length of the shadow according to that equation, we cannot manipulate the height of the pole by manipulating the length of the shadow.[30] Thus we supplement the previous diagram and equation with a model of the causal relations among these variables:

[30] This claim could be made more precise. There are of course changes that affect both the length of the shadow and the length of the pole. Woodward therefore gives a very precise definition of "intervention," such that an intervention on a variable X with respect to a variable Y involves a) setting the value of X independently of its other causal antecedents and b) doing so in a way that does not affect the value of Y except by way of X. For a more detailed discussion, see Woodward 2003 pp. 94ff., esp. pp. 98ff.

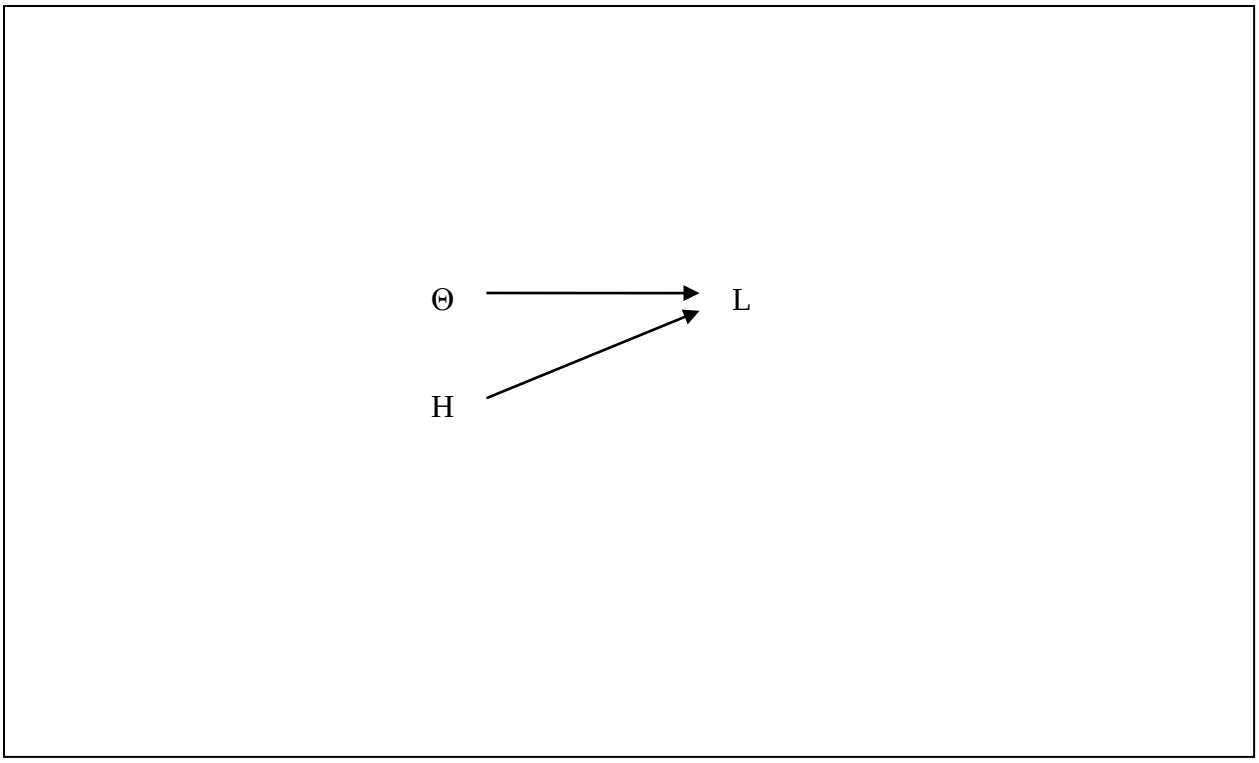

Figure 2

As noted above, all of these are variables ranging over states, not events. Indeed, this example involves no mention of time.[31] The causal relations here specified are time-neutral relations among states: this diagram does not represent a temporal order of events. What is essential is that certain asymmetric relations of causal explanation hold among the variables in question: by manipulating some of them, it is possible to manipulate others, but not vice-versa. And, on Woodward's view, this suffices to establish that we have genuine causation here. There is no need to discover some sense in which this particular causal model is entailed by or in some more nebulous way "backed by" a law of nature.

1.2.3: From causation to dependence

I've said that Woodward offers a detailed account of causation which promises to bear fruit in the philosophy of action. But causation is not the only dependence-relation that figures in practical thought. Consider, for instance, an

[31] While it is true that light moves at a finite speed, it need not have, and for most practical purposes involving flagpoles and shadows, light moves infinitely fast.

example from Anton Ford: "I am eating broccoli in order to eat some vegetables."[32] But there is not a causal connection between a comestible's being broccoli and its being a vegetable. So instrumental reasoning can grasp relations of non-causal dependence. One might also (and this is more controversial) wonder whether the relation between thought and action is always a causal one. Anscombe's examples of promising, marriage, and contracting seem to suggest as much: if these are the sorts of things that *cannot* be done accidentally, then there is some kind of non-causal relationship between thought and action in such cases.[33] And these are not just a few recherché cases; rather, conventional and linguistic acts generally involve non-causal relations.[34] So our account of grounding must include how agents understand non-causal dependences.

One might be wary of moving from a notion of specifically causal dependence to a more general notion of dependence. Pearl, Spirtes, Glymour, and Scheines have provided a detailed epistemology for causal claims. Woodward has provided an account of the semantics of such claims. And Gopnik's research suggests that these ideas are "psychologically real" – that they account not for just an abstract philosophical structure, but in some sense describe the ways human beings actually reason about causation. So the manipulationist account of causation has significant foundations. None of this can be said for the notion of non-causal dependence; indeed, until recently, analytic philosophers abjured such ill-defined metaphysical notions.[35] All the same, Fine's argument seems plausible: while Socrates exists iff {Socrates} exists, it seems overwhelmingly plausible the former is somehow explanatorily prior: {Socrates} depends on Socrates, but not vice versa.[36] In general, the notion of non-causal dependence, explanation, grounding – call it what you will – seems essential.

---

[32] Ford (unpublished) §7. I've changed the vegetable in the example from spinach to broccoli.
[33] Anscombe 1969 p. 61. But why can't these connections be causal? Woodward (2014) makes this point forcefully. The basic problem is that treating necessary or conceptual connections as causal connections leads to seriously flawed reasoning. To take Woodward's example, imagine that we are investigating the influence of high-density and low-density cholesterol on heart disease (2014 §6). Quite obviously, we might consider a third variable: total cholesterol. But if we treat this as depending causally on high and low density cholesterol, then we are forced to condition on it when determining whether low-density cholesterol has any direct effect on heart disease. But if we do this, the necessary connections among these three variables mean that any variation in low-density cholesterol will be accompanied by variation in high-density cholesterol (since we are keeping total cholesterol fixed). This quite clearly will prevent us from accurately measuring the effects of low density cholesterol. So we cannot treat relations of non-causal dependence as if they are relations of causal dependence.
[34] Thanks to Suzanne Love for emphasizing this point.
[35] Though see Schaffer 2015 for an interesting recent attempt to extend the causal model approach from causation to grounding.
[36] Fine 1994.

Thus, I want to reformulate the three essential theses I extracted from Woodward's account in terms of non-causal dependence. I said:

1: Causal knowledge is knowledge of dependences among variables of various sorts.
2: Knowledge of invariances is causal knowledge.
3: Shallow causal knowledge is possible.

The first thesis has a straightforward interpretation: variables of various sorts can depend on one another. For instance, being married (a legal and perhaps normative state) depends on getting married (a public event) which in turn depends on knowing one is getting married (a mental state). The second is more difficult, since Woodward characterizes invariance in terms of intervention, and the notion of intervention is not well-defined for variables linked by non-causal dependences. That said, invariance is in Woodward's system a replacement for law of nature. And there seem to be analogous concepts in the realm of non-causal dependence. A law of nature would state, in a universal and exceptionless manner, how one variable depends on another: for instance, in *exactly* what circumstances making a promise produces an obligation. An invariance, by contrast, need not be exceptionless and universal, but rather holds in at least some range of circumstances. Thus, we can say: making a promise non-causally produces an obligation - but there are a variety of circumstances, the precise details of which it is surprisingly difficult to spell out, under which making a promise does not produce an obligation. But in at least some circumstances, the relation between promising and obligation is invariant: a promise to do A non-causally grounds an obligation to do A, for a range of substitutions for A. Agents can know such an invariance holds without knowing the deepest natures of promises and fidelity. And this explains (3), that shallow knowledge of dependences is possible. Thus, I conclude that the pertinent features of Woodward's account of causation can be extended to an account of non-causal dependence.

1.3: REPRESENTATION AND SELF-REPRESENTATION

I began with this formula: practical thought represents itself as the ground of its object. In the previous section, I explicated the relevant notion of ground. In this section, I will explicate the relevant notion of self-representation. In §1.3.1, I discuss the general picture of representation I work with. In §1.3.2, I turn to self-representations, i.e. de se representations. In §1.3.3, I respond to some general worries about this picture of self-representation.

1.3.1: Representation

I said that intentions, like beliefs, are representational states.[37] What does this entail? Consider the following taxonomy of the mind: some mental states are non-representational (depression), others are representational (wishes). Of the representational states, some are not governed by an accuracy norm (imaginings), while others are (beliefs and intentions). Such states have both a sense and a reference – that is, they have both a characteristic inferential role and satisfaction-conditions – i.e. both an internal-functional role and an external standard. None of this amounts to cognitivism about intentions, but is a very generic taxonomy shared by a wide variety of theories. So long as you hold that an intention to do A may be satisfied or not satisfied insofar as it is executed successfully or not, so long as you hold that an intention to do A plays a characteristic role in reasoning, then you – in my sense at least – admit that intentions are representational states. But this rather minimal thought can easily be misunderstood, if the bare idea of representation is overladen with other, unnecessary features.

First, representations are not essentially conscious, or attended-to, or "present to the mind" in whatever sense we choose to give that phrase. While anyone with an imagination must know that we have a Cartesian Theater, which plays many and varied scenes, beliefs or intentions are not to be identified with goings-on in such a theater. I might, at a party, in a strange mood, be suddenly struck by the thought "M has a monster behind their eyes," without in any way *believing* that M has a monster behind their eyes; the sentence "2+2=5" might resound rather insistently in that theater, despite my total lack of inclination to believe such a ridiculous thing. And the same is true of intentions: as Anscombe says,

> "Or suppose I feel an upsurge of spite against someone and destroy a message he has received so that he shall miss an appointment. If I describe this by saying 'I wanted to make him miss that appointment', this does not necessarily mean that I had the thought 'If I do this, he will...' and that affected me with a desire of bringing it about, which led up to my doing so. This may have happened, but need not. It could be that all that happened was this: I read the message, had the thought 'That unspeakable man!' with feelings of hatred, tore the message up, and laughed."[38]

Nothing about the view that I offer requires that every intention is accompanied by some ringing pronouncement in the agent's Cartesian Theater along the lines of "I shall destroy his message out of spite!"

In addition to this "mind-mind" problem, there are some mind-body problems that have afflicted the idea of a representational state. For instance, we frequently specify the content of an agent's representation using English

[37] Why do I say "representational state" rather than "propositional attitude"? Because, like Lewis (1979a), I think that the objects of representational states are properties, not propositions. More on this odd thesis in a moment.
[38] Anscombe 1957 p. 17

sentences such as "L suspects that we are no longer friends." This has led some to the thought that we must find a sentence-like thing in the brain to which that belief corresponds. But perhaps connectionist models of the brain undermine that possibility.[39] One might also attack the idea of representational states from a different angle: Jaegwon Kim argued that because they are physically realized somehow, they must be causally irrelevant.[40] In response to these sorts of mind-body problems, I will do little more than kick a rock: claims like "N went in there [gesturing to an ill-labeled storage closet] because she thought it was a bathroom," or "M is sobbing because he has just heard the news" are, I take it, on about as secure a footing as any causal explanation whatsoever. And I leave it to the specialists to work out the metaphysical details; in the meantime, I simply presume that human beings can have representational states, and that these representational states can be causally efficacious.

1.3.2: Self-representation

Such theses about representation I simply presume; now let's turn to *self*-representation. Here is a summary of what I wish to show: it is widely accepted that first-person indexical beliefs like "I am Rudolf Lingens" have a special sort of content. Such beliefs attribute a property to the believer; they are true just in case the believer has that property. I propose extending this idea to allow that individual representations may attribute a property *to themselves*; they are true just in case *that representation* has the property. This suggestion may arouse worries about circularity and paradox, but it should not. We already accept that we may attribute properties to the representational states of others: we may say that the explanation of John's sudden departure was his (false) belief that his train was about to leave. Here, we attribute the property of causing a departure to his belief. All I propose is that some representations may attribute such properties (especially causal and grounding properties) to themselves. Clearly enough, this framework allows for paradoxical or puzzling states that attribute properties like "is false" or "is green" to themselves. But that's just a feature of our framework for talking about beliefs in general – it already allows for paradoxical states, such as "Everything I believe is false" or "Everything I believe is green." Introducing this new sort of self-referential representation introduces no new paradoxes or circularities, and – as I will argue over the coming chapters – is needed to make sense of the specific representational content of practical thought.

[39] cf. Garson 2015.
[40] Kim 1998; Woodward 2014 contains an extended response.

Now to showing it. It has long been noted that first-person beliefs, such as "I am on Fifth Avenue" have two important features: first, it's difficult to reduce them to non-indexical beliefs (e.g. "JSH is on Fifth Avenue"), and second, they play a central role in explaining action.[41] John Perry opens his classic "The Problem of the Essential Indexical" with this story:

> "I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch.[42]

Perry truly believed that someone in the market was making a mess; he might even have seen himself in a mirror, and thought "That man is making a mess," if he didn't realize that he himself was the man he saw. Until Perry realizes "*I* am making a mess," he won't stop to fix his leaking sugar. David Lewis considers the similar case of Rudolf Lingens, an amnesiac lost in the Stanford library. Dejected, he picks a book at random and begins reading; it is in fact a combined biography of Rudolf Lingens and map of the Stanford library. And so eventually Lingens comes to have detailed knowledge of the man that he is and the place where he stands; he may even know "Rudolf Lingens is lost in the Stanford library (but if he would just open the door labeled "Staff Only," he would soon find his way to the outside world)." The point is that it is possible to represent everything there is to represent about oneself, but so long as one does this in a merely third-personal way, there is still some essential class of representations one is missing: the self-representations. Lingens will only escape the library when he realizes "I am Rudolf Lingens, and this is the Stanford Library."

So when is a self-representation true? A first answer: if X believes "I am P," then that belief is true just in case X is P. But there are difficulties: sometimes we need a more specific "self" in our self-representations. Lewis writes,

> "Consider the insomniac. Tonight, as most nights, he lies awake for hours. While he lies awake, his state of mind changes little. He does not keep track of the time. So all through the night he wonders what time it is. To wonder is to lack knowledge, but what relevant knowledge does he lack? Not propositional knowledge; he knows, well enough, what sort of world is his. And not self-ascription of properties to his continuant self; he knows, well enough, what locus through space and time he occupies and what his various segments are like. He knows, for instance, that he spends the night of 13-14 February 1978 lying awake and wondering what time it is. To understand how he wonders, we must recognize that it is time-slices of him that do the wondering. A slice of the insomniac may locate the whole of the insomniac well enough in logical space and space and time. Yet that slice may fail to locate itself in space, in time, and in the population of slices of the well-located continuant insomniac."[43]

---

[41] e.g. Castañeda 1966, Anscombe 1975, Perry 1979, Lewis 1979a.
[42] Perry 1979 p. 3
[43] Lewis 1979a p. 527

The point here is that if we take "X" to refer to the entire continuing person, then we can't make sense of temporal self-location. So we should take "X" to be a time-slice of a person: if X believes "I am P" at t, then that is true just in case X is P at t.

There are a number of proposals about the underlying semantics of first-person beliefs. Lewis suggests that the objects of representations are not propositions, but properties. Many who mistakenly take themselves to be following Lewis, but instead follow a suggestion due to Quine, claim that the objects of representations are sets of "centered" possible worlds.[44] And Kripke suggests that "I" is a particularly strange referring term. But the points I wish to make now do not depend on this, as each theory largely agrees on the surface details. On any of these theories, a self-representation has two moving parts: a predicate and a "self." The "self" can be a person, a time-slice of a person, or, as I argue, a particular representation. The self-representation is true just in case the predicate holds of that self. Putting this in terms of centered possible worlds, we start with a center (a person, a time-slice, a representation) and the find the set of worlds where the predicate holds of that object: and thus a set of centered possible worlds.[45] Thus, throughout my dissertation, when I need to specify the content of a self-representation, I will do so by specifying a predicate, and this often by means of an open sentence with one free variable. This content can be used to construct a sentence with the appropriate self-referring term, or a set of centered worlds, or simply a self-representation, as the reader prefers.

I now want to suggest that we should countenance not just time-slice-centered representations, but *representation-centered* representations. Here's an overview of Lewis' argument strategy: first, imagine that beliefs self-locate in merely logical space, in the space of possible worlds. But it seems that a person might know exactly where they are in logical space – everything about their world – yet not know *where* in their world they are. They have a perfect map, except that it's missing a "You Are Here" marker. So, Lewis suggests, we need beliefs that locate themselves in the space of persons in a possible world. But it seems a person might know everything there is to know about themselves, except what time it is – where they are in time. So we need beliefs that locate themselves

[44] It is a deep mystery to me why anyone cites Lewis as the author of the idea of centered worlds. In his 1979a Lewis does not bury the lede: on the second page, he claims clearly that the objects of beliefs are properties, a claim he repeats in 1983 (p. 29). He mentions the idea of centered worlds only in passing, clearly attributing it to Quine in both cases (1979a pp. 531-533; 1983 p. 25 n. 18).

[45] Lewis suggests exactly this procedure for translating his theory into Quine's terms (1979 p. 532).

in time as well. What I will argue in the following chapters is that we must also allow for even more specific self-location to account for the representational nature of intentions. In particular, we need representations which locate *that very representation*.

The argument for such representations will come later. For the moment, I simply want to discuss what such representations would be like, given that they exist. I'll start by contrasting the inferential behavior of representation-centered representations with that of person-centered representations. Consider the following inference involving person-centered representations:

> RS believes "I am RS"
> RS believes "RS is P"
> On that basis, RS believes "I am P"

This is, obviously, a perfectly valid inference. Now consider the same form of inference,[46] but with representation-centered representations:

> RS believes "This very belief is P"
> RS believes "That belief is Q"
> On that basis, RS believes "This very belief is Q"

This inference, on the other hand, is quite invalid.[47] And the reason is obvious: the phrases "this very belief" in the premise and conclusion refer to different beliefs. This inferential fact reflects a basic *logical* point about the truth-conditions of self-representations. Just as we cannot ask "is "I am RS" true?" without specifying which agent we are talking about, we cannot ask "is "This very belief is P" true?" without already specifying which particular representation we're talking about.

Why aren't representation-centered representations self-validating? Consider the contrast with "This proposition is true," an ungrounded and perfectly circular proposition, or perhaps worse "This proposition is either true or false," which seems to be fairly self-validating. At the end of the day, we have no real way of saying whether the first proposition is true, or what the truth of the second would mean. The situation is quite different with representation-centered representations. The content of such representations is just some perfectly normal predicate. We already speak of the properties of beliefs and various mental states – for instance, we say "He was anxious because of the cassowary's unexpected appearance," and thereby attribute the property "caused by a cassowary" to his anxiety (and, notably, to *this particular instance* of anxiety, not to anxiety in general – that has more sources

[46] By "same form," I mean that the inference involves a) a de se belief, b) a belief about the object of that de se belief and c) a new de se belief.
[47] It turns out that this fact is essential to distinguishing between intended and side effects. I discuss this important fact in more detail in Chapter Two (§2.4).

than large, flightless, and extremely cranky birds); when we believe "She ran out the door because she thought it might be raining," we attribute the property of "causing running" to her belief. And we have a good grasp of how to assess the truth of these claims – how to determine whether there is any such mental state with such a property. In the case of representation-centered representations, we are faced with just the same question: does this mental state have a certain property? In this way, self-representations can be well-founded in ways that perfectly circular propositions are not. In the same vein, self-representations are not essentially self-validating: despite having unusual truth-conditions, they can still be quite false.

All in all, representation-centered representations ought to strike you as having some rather odd features. What I intend to argue over the course of this dissertation is that these are exactly the features needed to account for intentions and intentional action. In so doing, I hope to make representation-centered representations more palatable – by showing how they fit into familiar forms of thought, reasoning, behavior, etc.

1.3.3: Worries: circularity and nominalization

One might worry that representation-centered representations introduce an unacceptable kind of circularity into our theory of belief. In this section, I will argue that they do not introduce any new and vicious circularity: if there are any circularity problems regarding representation-centered representations, they are problems generally for theories of belief.

First, one might worry that such representations have unacceptably circular contents. I respond: de se representations are not about themselves in virtue of some feature of their content, e.g. the sort of complex construction used by Gödel to achieve self-reference in arithmetic. Rather, these representations are about themselves in virtue of their form. Their contents are just properties, of a perfectly ordinary sort. And they are true just in case they bear that property. Now, it is possible, with the right choice of property, to generate paradoxical de se beliefs: "is false," for instance. But this is true of beliefs in general: the right content in the right circumstances can easily turn paradoxical. For instance, if we admit that one thinker can think about another's thoughts, then we can easily generate a paradox: if Susie remarks, "Billy is speaking the truth," just as elsewhere Billy remarks "Susie is uttering a falsehood," then we representations with circular and paradoxical contents. In brief: representation-centered representations do not automatically have circular contents. While they may bear such contents, this is a feature of representations in general. So representation-centered representations introduce no new circularity here.

One might however worry that there is some deeper circularity afoot: that there is some circularity involved in specifying not the content but the truth-conditions of a representation-centered representation. Consider things in the centered-worlds framework: in order to specify the content of such a representation, we need to specify a set of centered worlds, i.e. a set of world-individual pairs. But the individuals in question will be representations with the very content in question: so in order to specify the individual, we already need a specification of its content – and now we are going in circles.[48] I respond: if this is a problem, it is a problem not just for de se beliefs, but in fact for any understanding of contents as sets of possible worlds. Consider: if the content of a belief is a set of possible worlds, then many of those worlds will include the very belief in question. So in order to specify one such world, we'll need to specify that belief, i.e. its content – and now we are going in circles. The way out, I think, is that specifying a set of possible worlds does not require specifying every detail of that set. Thus, consider the causal properties I am centrally interested in, such as "will cause a house to be built." This causal property specifies a set of centered worlds, i.e. a set of individual-world pairs such that at that world that individual will cause a house to be built. A representation-centered representation with that content is true just in case it is one of those individuals. And this specification of the content of that representation involves no obvious circularity. If there is some other, deeper, metaphysical circularity involved, then that circularity afflicts not just de se representations, but representations in general. So again, representation-centered representations introduce no new and worrying circularity into our account.[49]

Issues of circularity aside, one might worry that my use of nominalizations such as "belief," "intention," and "representation." One might say: really, these are just words we use to talk about people and their properties, e.g. facts like "John believes that Mack is faithful." But if self-reference is reference to a particular, then there is no particular to refer to: and therefore there cannot be representation-centered representations.[50] I respond: the premise is of course correct. While I will throughout the dissertation speak of beliefs and intentions, this is always just a convenient shorthand for talking about states of believing and intending: intentions are not objects on the order of

[48] Thanks to Adam Marushak for suggesting this issue.
[49] Thanks to James Shaw for several helpful discussions on these issues.
[50] Thanks to Kieran Setiya for suggesting this difficulty. Setiya discusses a similar problem for his theory (2007a p. 45 n. 38), where (as I read him) he distinguishes between token and type mental states, and suggests that self-referential mental states refer to their type, rather than their token. In my view, this is a mistake; a state type is something like "x believes that Apollo loves Starbuck," which type may be instanced in a number of different heads (Dee, Anders, etc.). If this is the case, state *types* are not the proper subject of self-reference: *states* are. If we wish, we can call a fact like "Dee believes that Apollo loves Starbuck" a state-token, but that invites the thought, evident in Setiya's footnote, that state-tokens are something like occurrences or events. So I prefer simply to talk of the properties of states.

oranges. But this does not mean that states of intending are not suitable centers. All that centering requires is that we should be able to talk about the properties of the putative center: and we can say, "John believes that Mack was faithful because of his abiding love." Or instead we might predicate "It is so sweet that," or any other thing which may be intelligibly said of a belief state.[51] And while perhaps there are some properties that are not intelligibly applied to states of belief ("is prime," perhaps?), there are nevertheless many which are, including the chief objects of my inquiry: causal and explanatory properties. So my use of convenient nominalizations does not introduce any insurmountable difficulties.

In summary: de se thoughts attribute a property to their thinker; we say: they are true just in case their thinker has that property. I am investigating the possibility of thoughts centered on themselves: they are true just in case they have that property. One might worry either that my reliance on nominalizations, or some hidden circularity, threatens the coherence of this proposal. But it does not. To put it formally:

> Where X ranges over thinkers and P over properties,
> R(X,P) means: X has a self-centered thought with the content P
> T(R(X,P)) means: the thought R(X,P) is correct
> T(R(X,P) iff P(R(X,P))

This formulation involves no problematic reliance on nominalizations, and speaks only of persons, properties, and facts; the properties in question are not circular, but are perfectly ordinary properties. While of course this formula breaks down under some substitutions for P, such as "is false," this reflects more on the difficulties of such properties than it does on this formula. Similarly, there are other properties which do not clearly apply to representational states at all, such as color properties. But again, this is nothing particular to do with de se representations: there's something fishy about wanting that someone else's thoughts be chartreuse. So while some properties do not clearly apply to representational states, others clearly do; of especial relevance to my project are causal properties. So, I conclude: there are no hidden incoherences in the idea of a representation-centered representation.

---

[51] One may raise difficulties, as George Wilson does, about the individuation of mental states (1989 pp. 278-280; this is the passage that Setiya is addressing in the aforementioned footnote), and therefore the difficulties of attributing causal powers to them. But while Wilson's objection is framed as a problem for self-reference theories of intention, it is in fact a problem for any theory which wishes to attribute causal powers to states, mental or otherwise. Yet, I take it, we intelligibly and correctly do so all the time: I cite, for instance, the mass of the table as a cause of the fact that it broke Tommy's foot when we dropped the damn thing. So while there may be difficulties in individuating states, I take it that these difficulties are surmountable. More theoretical support for this contention is, as I discussed in §1.2 of this chapter, provided by shifting from *DN* to Woodward's manipulationist account of causation.

1.4: THE ROAD AHEAD

I began with a problem: by focusing on particular aspects of intentional action, philosophers have been led to conflicting accounts of intentional action. I hope to remedy the situation by providing a unified account: practical thought represents itself as the ground of its object. In the next three chapters, I argue that this idea is both necessary and sufficient to account for each of the three aspects I identified. In Chapter Five, I consider several "big-picture" objections to this account, e.g. the charge that it requires too much conceptual sophistication of animals and small children. In Chapter Six, I return to the animating questions of action theory discussed earlier, and show how this account applies to first-order normative doctrines, the metaethical ambitions of constitutivism, and the philosophy of mind generally.

2.0: DEVIANCE AND OTHER PRACTICAL CATEGORIES

2.1: PRACTICAL CATEGORIES

Consider this narrative from Donald Davidson:

> "This morning I was awakened by the sound of someone practicing the violin. I dozed a bit, then got up, washed, shaved, dressed, and went downstairs, turning off a light in the hall as I passed. I poured myself some coffee, stumbling on the edge of the dining room rug, and spilled my coffee fumbling for the *New York Times*."[52]

Shaving, dressing, the darkness in the hallway: these are all *intentional*. Stumbling and spilling are not: they are *failures*. And while Davidson does not mention it, he also made some noise walking downstairs, but that was presumably not his goal, but rather a *side effect*. Call these various descriptions *practical categories*. One essential task for any theory of intentional action is to provide an account of these. I presume this much of a starting point: these categories represent various ways the agent's thought can relate to the world. For instance, part of what separates intentional actions from failures is that the latter do not match some aspect of the agent's thought – what you intended to do you didn't do. The difficulty therefore is to spell out what these various relations consist of.

We can begin our account of this suite of relations by considering Davidson's famed argument for causalism. In effect, Davidson argues that in order to distinguish intended from side effects, we must rely on causal notions. As he puts it,

> "...a person can have a reason for an action, and perform the action, and yet this reason not be the reason why he did it."[53]

To hold otherwise blurs the distinction between those things we intend, and those we desire but merely foresee. One might, for instance, make a hiring decision based purely on citation rates, know that this will lead to the hiring of one's close friend, and yet nevertheless intend to hire on merit rather than friendship. The thought about citation rates, unlike the thought about friendship, explains the hiring. Davidson therefore proposes that our thoughts about

[52] Davidson 1971 p. 43.
[53] Davidson 1963 p. 9.

intended effects are causally efficacious in ways that our thoughts about potential side effects are not. In retrospect, this seems hasty: Davidson's argument does not rule out forms of non-causal dependence, which might equally do the job of sorting out intended from side effects. Thus, I reinterpret Davidson's conclusion: intended effects *depend on* our motivating reasons, in ways that they do not depend on our thoughts about side effects. And this suggests a general scheme: practical categories represent various ways the world can depend on the agent's thought.

Unfortunately, both causal and non-causal dependence introduce a new problem: the notorious "deviant" or "wayward" causal chain, exemplified in this example of Davidson's:

> "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."[54]

The difficulty here is that the climber's reasons for murder are causally efficacious at bringing about the intended death, but not "in the right way." While Davidson himself eventually despaired of a non-trivial account of deviance, this is the conclusion of last resort.[55] If there is an account to be had, we would like to have it.

The task of this chapter is therefore to define four practical categories: intended, side effect, failure, and deviant.[56] But §2.2 addresses only an artificially restricted part of the topic. I consider only *causal* dependence, rather than dependence generally. And I consider only the problem of distinguishing deviant causation from intentional action, rather than the practical categories generally. Drawing on theories of causal understanding from the philosophy of science, I argue that agents *understand* how their thoughts connect to the world. When that understanding is correct, then thought and world meet non-deviantly. In §2.3, I respond to some objections to this proposal. In §2.4, I drop the earlier restrictions, and show how to use these ideas to define the practical categories generally. To do so, I rely on a surprising conclusion of the foregoing. Practical thought must be de se or self-locating in a particularly strong sense (discussed in Chapter One §1.3): the practical thought of agents locates itself in the causal structure of the world.

---

[54] Davidson 1973 p. 79. The example here is formulated in terms of causal dependence, but a theory of non-causal dependence may face the same problems, e.g. as Sarah Paul has argued (2010).
[55] Davidson 1978 p. 87.
[56] As it turns out, these categories are grammatically quite awkward. On my view, this is an artifact of the fact that they apply to a wide range of objects: processes, events, states, anything that may depend on an agent's practical reasoning. But one might well wonder whether this grammatical awkwardness is a sign of underlying logical diversity - i.e. that by defining these categories so broadly I am improperly lumping together logically disparate concepts which would be better kept separate. In this as in so many other things, I suggest that the proof is in the pudding: if I can give a unified account of these categories which applies to these various objects, then this is reason to conclude that the awkward grammar is just that: awkward grammar, nothing more.

2.2: DEVIANCE AND CAUSAL UNDERSTANDING

I suggest a simple theory:

> *Deviance:* The dependence of the agent's action on their thought is deviant if and only if it deviates from the agent's understanding of how their action depends on their thought.[57]

While this has long been thought to be a partial solution to the problem of deviance, it is almost universally believed to be merely partial. To show why *Deviance* has been so disdained, we need to consider the distinction between *basic* deviance and *non-basic* or *antecedent* deviance.[58] Basic deviance concerns basic actions; non-basic deviance concerns non-basic actions. A basic action is simply an action which is not done by means of any further action.[59] If X is doing B basically, then there is no A such that X is doing A in order to do B. The classic cases of basic actions are bodily movements: when I intentionally raise my arm in the normal way, there is nothing else that I am doing intentionally in order to raise my arm.

Given that definition, almost all authors have accepted that *Deviance* suffices to explain non-basic deviance. But it fails to account for basic deviance. Let's consider a pair of examples.

> "A man may try to kill someone by shooting at him. ... [T]he killer misses his victim by a mile, but the shot stampedes a herd of wild pigs that trample the intended victim to death."[60]

This is a case of non-basic deviance: the killer intends to kill his victim by shooting him. While he fires his gun in the normal way, the connection between the firing and the killing is deviant. Contrast this with a case of basic

---

[57] This proposal is heavily indebted to an idea independently developed by Nomy Arpaly and Ralph Wedgwood: that non-deviance requires that an intention cause an action in virtue of rationalizing that action (Arpaly 2006 p. 69, Wedgwood 2006 p. 670). Unlike other definitions, this proposal begins to involve the *content* of the agent's thought, rather than merely treating intentions like another causal variable. Unfortunately, I reject this definition for two main reasons. First, I am uncertain whether there is a notion of rationalization that can do the appropriate work. It cannot be mere justification, since merely intending to do A is frequently no justification at all for doing A (cf. Broome 1999 p. 410). But if not justification, then what sense of rationalization do Arpaly and Wedgwood have in mind? Second, on this view, rationalizing features are just parts of the causal order. And causation can be transitive. So it must be possible for an intention to cause, in virtue of its rationalizing features, some unusual subpersonal event, which in turns causes an action: and thus the intention would cause, in virtue of rationalization, an action. Yet this seems deviant. So I suspect that Arpaly and Wedgwood's proposal does not work (though I do not have the space to develop these criticisms in detail here). Nevertheless, my own solution was prompted by thinking through theirs, and I am thus deeply indebted to their work.

[58] Alfred Mele suggests a third category, which he calls "tertiary" deviance (1992 pp. 207-208). While Mele's specific case is somewhat complicated, the basic problem seems to involve acting on the basis of an accidentally true or Gettiered belief. Thus "tertiary" deviance seems to be a primarily epistemological rather than practical notion; I therefore set it aside for the moment.

[59] There is some debate on whether there are any basic actions (cf. Thompson 2008 Chapter 7; Setiya 2012 §1; Lavin 2013). I do not wish to take a stand on this issue; I introduce their definition not on the supposition that anything answers to it, nor on the supposition that it is a philosophically important category, but rather to demonstrate that, even if they exist, they require no special treatment in the philosophy of action.

[60] Davidson 1973 p. 78; Davidson attributes the example to Daniel Bennett.

deviance discussed earlier: Davidson's murderous mountain climber. In that case, the climber does not let go of the rope by doing anything else; the deviant connection therefore lies not between two actions, but between an action and an intention. As I said, it has seemed plausible to many that *Deviance* can account for the former case. The agent expected the firing to cause the killing via a gunshot wound, not via pigs. So the connection is deviant. But it has seemed to many that no analogous story is possible in the latter case, since Davidson's climber does not have any causal understanding of how his intentions cause his actions.

I hold that this conclusion is wrong, that it both overestimates the sophistication necessary to understand the causation of basic action and underestimates how sophisticated ordinary agents actually are: the causal understanding of ordinary agents is sufficient to define both antecedent and basic deviance. But showing this will require a detour into general questions about causation and causal understanding that are quite independent of action theory. For the traditional assessment of *Deviance* seems to combine two hazy theses: that causal understanding must approximate scientific understanding, and that scientific understanding consists of knowledge of a) chains of events and b) the natural laws connecting those events. Applying these ideas to Davidson's climber, we get the verdict that if the climber understands how their intentions cause their actions, that would be the sort of knowledge a physiologist might have of that connection, which would consist of knowledge of a) the neural and muscular events intervening between the brain and the hand's movement and b) the natural laws which connect such events.

But these are not action-theoretic theses; they are theses about causation and causal understanding. Therefore in order to assess whether *Deviance* is adequate to the problem of basic deviance, we need to turn from the philosophy of action to the philosophy of science. So how do these ideas fare as general models of causal understanding? At one point, they were widely accepted in the form of the "deductive-nomological" model of explanation, due to Carl Hempel and Paul Oppenheim.[61] But natural laws and perfectly basic events turn out to be hard to come by. For instance, given the conflict between general relativity and quantum mechanics, even our current best physics does not provide any exceptionless natural laws. Yet nevertheless there must be some clear sense in which our current physics offers us a better causal understanding of celestial mechanics than Ptolemy possessed; in general, we wish to maintain that science progresses. Similarly, we wish to hold that the person who believes that bleach removes stains has a better understanding of bleach's causal properties than the one who believes that bleach makes soups savory. Thus the deductive-nomological account of explanation is typically

[61] Hempel and Oppenheim 1948 contains the classic statement of the view.

supplemented by a different model of causal understanding.[62] On this view, agents can understand how A causes B, despite not knowing how to reduce A and B to fundamental physics (or anything else) or which natural law applies to the reducing events. Causal understanding instead consists of some information which narrows down which fundamental causal laws apply to the situation.

I mention this not because I endorse the deductive-nomological model - I don't - but to emphasize that even the most austere definitions of causation wind up with significantly more liberal definitions of causal understanding. Less austere definitions, such as the "interventionist" or "manipulationist" model defended by James Woodward (and many others; I discuss this model of causation in more detail in Chapter One §2.2),[63] give a more straightforward account of causal understanding. Woodward writes,

> "We have at least the beginnings of [a causal] explanation when we have identified factors or conditions such that manipulations or changes in those factors or conditions will produce changes in the outcome being explained."[64]

On this view, causal understanding just consists of true beliefs about causation, such as "If I depress the gas pedal in such-and-such circumstances, my car will go forward."[65] On the deductive-nomological view, this is not a true causal claim, but it is an accurate causal understanding, since it narrows down which fundamental causal claims apply to the situation. What I hope to do is prescind from the question of which of these is the correct theory of causation, and focus on the relatively common ground of causal understanding.

Given that, we can ask: what does ordinary causal understanding about action consist of? We can approach this by asking: what is the surface causal structure of action? A first answer: action, some worldly goings-on, is caused by a mental state, intention. Action depends on, is explained by, intention. But action does not depend on only intention; if it did, we would always succeed at doing what we tried to do. Instead, agents often enough fail to achieve their goals. So action also depends on external factors – the agent's circumstances. To take a simple example, I won't move my arm unless a) I intend to move my arm and b) my arm is not tied down.

What I wish to argue is that agents in general have extensive such knowledge – knowledge of how their actions depend on their intentions and their circumstances. This is because such knowledge is essential to being an effective agent. In particular, it is an essential component of means-end, instrumental reasoning. It plays two main

---

[62] For instance, Hempel 1965 p. 361. Notably, the same basic idea is present in a wide variety of other accounts of causation (e.g. Railton 1981 p. 240, Lewis 1986 pp. 214-221, Kitcher 1989 p. 414).
[63] e.g. Pearl 2000; Spirtes, Glymour and Scheines 2001; Gopnik and Wellman 2012.
[64] Woodward 2003 p. 10
[65] The example is from Woodward 2003 p. 18.

roles: first, explaining why agents don't try to do the things they can't do in present circumstances, and second, explaining why agents alter their circumstances to enable them to do other things. To return to the initial Davidsonian narrative: why did Davidson first get up, then walk to the bathroom, then shave? Because he knew that if he tried to walk to the bathroom while lying in bed, he would not walk to the bathroom. Why does he try (successfully) to walk to the bathroom after he gets up? Because he knows that in such circumstances, if he tries to walk to the bathroom, that will put him in the right position to shave. These claims are rather banal, but that's the point: this sort of knowledge is both essential and ubiquitous.

One might however resist these ideas for a pair of reasons. First, one might hold that instrumental reasoning only involves thinking about an action and its effects, not the causal relations between actions, intentions, and circumstances. This response traces at least to Davidson, who held that the "primary reason" which was the cause of an action was a belief and a desire; the belief was of the form "[action] A under the description d has property [P]."[66] This suggests[67] that instrumental reasoning turns on beliefs about actions (or action-types) and their desirable intrinsic properties or downstream effects. I suggest that this claim is due to a focus on successful actions: when an agent acts successfully, we have an action with various properties fit to be the subject of the agent's belief. But we also need to explain why the agent didn't attempt all those other actions which were impossible in present circumstances yet which, if successfully performed, would also have had those properties. And in order to explain *that*, we need to have recourse to the agent's beliefs about the efficacy of intentions to perform those actions in their current circumstances. Similarly, when we look at the downstream effects of actions, some of the most relevant include altering the agent's circumstances to enable them to perform others. But we can't explain that if the agent has no thoughts about the circumstances in which trying to perform those later actions will or won't be successful.

So much for the first objection. But the second is more serious. My account implies that agents in general have some concept of intention. But many hold that while animals and small children act, they have no concept of intention. So they can't understand how their intentions cause their actions. While I address this objection in more detail in Chapter Five (§5.3), I want to here suggest why I think it is misguided. But first I must admit that the objection is, in a sense, quite right: there are a wide variety of animals that act yet which lack such knowledge. To take a familiar example, consider the egg-laying activities of *Sphex ichneumoneus*, a digger wasp. The wasp's egg-

[66] Davidson 1963 pp. 4-5.
[67] It merely suggests, since the property P could be just anything, including a property like "Depends on my intentions and my circumstances." But Davidson's actual examples are uniformly (cf. Davidson 1963, 1969, 1971, 1973, 1978) about the intrinsic properties or consequences of an action, not its causal antecedents.

laying behavior is complicated but extremely routinized: any interference, even if it is extraordinarily trivial, causes the wasp to go back to the beginning of its routine and start from the top.[68] The wasp's behavior thus presents the appearance of intentional action, but its total inability to engage in the most basic flexible instrumental reasoning suggests that this behavior is instead a simple stimulus-response program.[69]

So I'll happily admit that *Sphex* wasps lack causal knowledge, but they also don't engage in intentional action. A more serious challenge comes from more cognitively sophisticated animals (housecats, perhaps) - animals which are intentional agents, i.e. which get around in the world by means of representational states such as beliefs and intentions and reasoning on the basis of those states. Such animals engage in intentional action and instrumental reasoning, but – the objection runs – lack knowledge of the causal properties of intentions, since they lack the concept of intention. This is because, for instance, they fail the "false belief" task: they cannot accurately predict the behavior of other agents with false beliefs.[70] But it turns out that the import of the false belief task is widely misunderstood. Rather than marking a sharp divide between agents with a "theory of mind" and agents without, it represents one step along a long series of developments of that theory.[71] For instance, some animals which fail the false belief task nevertheless pass the "knowledge-ignorance" task, which requires them to predict the behavior of other animals which are ignorant of various facts.[72] This suggests that such animals have at least some concepts of mental states.

Indeed, we should in general be suspicious of theories of content on which one particular test is criterial for possession of a complex concept such as "intention." Consider addition. Most elementary school students know that 2+2=4 (some even know that 25+26=51). So they have the concept of addition. But this does not entail that they are capable of applying the concept of addition everywhere it applies - for instance, to large numbers, or negative numbers, or complex numbers, or infinite series. They can apply the concept to a part of its overall domain.[73] Similarly, I suggest that agents which engage in instrumental reasoning are at least capable of applying

[68] Wooldridge 1963 p. 82. The topic of "sphexishness" was taken up in philosophy by Douglass Hofstadter (1982) and Daniel Dennett (1996) in connection with free will. I wish to avoid that subject, and instead focus on the cognitive implications of the wasp's behavior.
[69] I don't mean to suggest that the line between stimulus-response agents and intentional agents is everywhere sharp or well-defined. But as Anscombe reminded us (1961 p. 59), the fact that some cases are hard does not imply that every case is hard.
[70] See Carruthers 2008 for a powerful development of this common claim.
[71] Wellman and Liu 2004
[72] Kaminski et al. 2008
[73] See Wilson 2006 for a development of this conception of concepts, and Li (unpublished) for its application to the case of addition.

the concept of intention to themselves, even if they do not apply it to others or to instances of false beliefs. And this is simply because, as I've argued, effective instrumental reasoning requires knowledge of how one's own actions depend on intentions and circumstances.

So I conclude that intentional agents typically have knowledge of how their actions depend on their intentions and circumstances. How does this solve the problem of deviance? Consider Davidson's murderous climber. Clearly enough, the climber has no idea which events intervene between his intentions and the unclenching of his fists. And he has no idea about any natural laws – or even physiological laws – which apply to his situation. But he clearly does have some causal understanding of his situation. For instance, he does not take his actions to depend on his nerves. If he did, he would take more care to ensure that his hands do not tremble at an inopportune moment – say, a time when their unclenching would lead to *his* death. Thus he (falsely) takes his nerves to be causally irrelevant to his grip. So when his nerves do cause his grip to loosen, that deviates from his causal understanding. This, then, is how *Deviance* applies to basic actions: the agent's causal understanding is not found in knowledge of intervening events or natural laws, but in knowledge of patterns of dependence among actions, intentions, and circumstances. In cases of deviance, the way the agent is getting around in the world diverges from the way the agent takes themselves to be getting around in the world, but through a stroke of luck, they achieve their ends anyways.

We can contrast this approach to deviance with other proposed definitions. Consider a classic theory, most forcefully articulated by Harry Frankfurt. According to Frankfurt, non-deviant causation is *immediate guiding* causation.[74] So let us consider a case where there is neither guidance nor immediacy between intention and action, for instance, a paraplegic neurally connected to a robotic prosthesis. The connection between their intention and the movement of the limb is thoroughly mediate; it involves complicated computers in other rooms, wireless signals, technicians monitoring the limb's performance, and so on. And let us say that there is, in this instance, no guidance between the intention and the action. Perhaps the agent has had some time to grow used to moving the limb,

[74] Frankfurt 1978 p. 158; Frankfurt is followed by a wide array of other authors (Brand 1984 Part IV, Thalberg 1984, Bishop 1989 pp. 167-72, Heckhausen and Beckman 1990, Mele 1992, Mele and Moser 1994 p. 46, Setiya 2007a p. 32, Setiya 2012 pp. 293-294). I suspect there are also some problems with this definition other than the one I discuss in the text. For instance, surely there are some mediating events between human intentions and actions: so we are owed an account of the right kind of mediation – and one that would apply to any possible intentional agent, not merely human beings in ordinary circumstances. And since guidance is just a generic causal notion (one magnet may guide its course towards another), then surely there is such a thing as deviant guidance (say, if an intention had some curious magnetic properties): so we are owed an account of the right kind of guidance. But set these problems aside.

clenching the "hand," etc., but in this case receives no perceptual feedback (visual or otherwise) about the limb's movement. Without feedback, there can be no guidance.[75] Yet all the same: when the agent decides to clench their "hand," their "hand" clenches: and this, I claim, is clearly intentional, even if both mediate and unguided. *Deviance* allows for this, since the agent has some understanding of the circumstances under which trying to clench their robo-fist results in a successful clenching, and this is just such a circumstance.

Thus *Deviance*, unlike other proposed definitions of deviance, allows that agents can reason along any available pathway from intention to action, so long as that pathway is known to them. Put another way, when I discover that a manipulative neuroscientist has routed my neural impulses through an orbiting satellite, that does not forever prohibit me from successfully acting intentionally; it just means I need to be careful around parking garages (the concrete blocks the radio waves). This illustrates how *Deviance* would apply to the many and varied cases of deviance I have not mentioned. In such examples, the agent is typically unaware of the novel connection between their intentions and their actions: thus *Deviance* manages to simultaneously capture the deviant nature of these cases and the non-deviance of the paraplegic above.

2.3: SOME OBJECTIONS

The problem of deviance puts serious strain on definitions of intentional action; traditional definitions of deviance have borne this load directly. But *Deviance* instead relies on the individual agent's causal understanding to do a great deal of work. Thus *Deviance* faces challenges in cases where the agent's causal conception is unusual in one way or another. In this section, I work through a number of such challenges, and in so doing show how *Deviance* applies to a wide variety of cases of both intentional action and deviant causation.

2.3.1: The undergraduate occasionalist

Consider a student who, on encountering the marvelous doctrines of Nicolas Malebranche, embraces occasionalism wholeheartedly (but perhaps ineptly). "My thoughts and intentions are causally inefficacious!" they declare. And

[75] Guidance is not a non-causal notion (pace Frankfurt 1978 p. 158), nor a peculiar sort of a glow certain causal connections take on: it is a structural property of certain causal systems, such that deviations in the value of one variable from a set value cause that variable to return to the set value. To take a mundane example, the causal structure of most flush toilets involves guidance of the water level in the tank by a ballcock and a flapper valve. In general, guidance of one variable requires that that variable "feed back" into its causes: this is just what it is for a system to be disposed to return that variable to a set value.

yet they go around writing papers (intentionally), drinking coffee (intentionally), catching the bus (intentionally), and so forth. How can *Deviance* capture this? Their intentions cause their actions, yet they believe that this cannot happen. So the connection between their practical thought and the world diverges from their conception of that connection. So *Deviance* (wrongly) declares that all their actions are deviant rather than intentional.

Or so one might argue. I respond: in addition to their occasionalist doctrine, such an agent has clearly also got a perfectly normal causal understanding of the world, seeing as how they get around it in a perfectly normal way. They understand that they can't walk through walls if they try, but that they must first turn the knob, open the door, walk down the stairs, etc. if they wish to get anywhere. And this causal conception is the one that their means-end reasoning relies on. So it is this conception that we should look to when determining whether their actions are deviant. In general, when agents have multiple or conflicting causal conceptions, *Deviance* requires that we look to the causal conception which is operative in their means-end reasoning.

2.3.2: Brunel and the Box Tunnel

Consider Victorian engineer Isambard Kingdom Brunel, building a railroad according to the precepts of classical mechanics. When the Box Tunnel did not collapse on the first train to pass through it, was that mere luck? Does *Deviance* entail that this was just deviant or accidental success? I suggest that the earlier points concerning causal understanding again prove fruitful. Whether or not the theorems of classical mechanics are genuine causal laws, knowledge of classical mechanics is genuine causal understanding. While it is not as deep (perhaps) as knowledge of quantum field theory, it is not therefore false or incorrect in the way that a belief in the efficacy of homeopathic remedies is. Put another way: classical mechanics is in fact the best available theory of the motion of medium-sized dry goods. Therefore, while Brunel might have lacked knowledge of the fundamental natural laws, my account does not require that such knowledge be present for accurate causal understanding. And the same point applies to agents going about their everyday business: even if they do not rely on natural law to make their way around, they do rely on accurate causal understanding when they successfully act.

2.3.3: Bathtubs and blessed water

One might worry that the claims of the previous section let rather too much in. Don't they show that anyone who succeeds has *some* causal understanding? And isn't deviance thus impossible? Consider therefore a ye-olde-tyme

community in which it is widely believed that blessed water prevents wounds from becoming diseased. As it happens, their belief is false: whether water is blessed never makes a difference to infection. On the other hand, their belief is also quite reliable, since the bathtubs they use contain an antiseptic chemical which leeches into the water. This belief about the efficacy of blessed water clearly plays a role in their means-end reasoning: when someone is wounded, they always make sure to get a priest to bless the water before bathing, and when a tub is not available, they will pour blessed water over wounds (to no effect). So their causal conception is false but usually (though not always) reliable. So when they succeed at preventing infection, this is in part due to luck.

This example reveals that a belief's reliability and its truth can sometimes come apart. Therefore, accidental success (another way of saying "deviance") comes in degrees. One may be accidentally successful because of a freak occurrence, as is typical in traditional examples of deviant causation, or because of a reliable but not strictly true belief, as in the case above. I hold that both of these are accidental successes: but some accidental successes are more accidental than others. Among those causal conceptions which are properly false, there is a scale of more and less reliable, and therefore a scale of more and less accidental success in action.

2.3.4: Mabel in a crate

Next consider cases where agents have no causal understanding, or only extremely minimal such understanding. Take a relatively unsophisticated person who intentionally blinks: they have no idea about how their intentions cause their blinking.[76] It's worth noting that the supposition here is quite extreme: the unsophisticated agent does not even realize that taping their eyelids open will prevent them from blinking. All the agent knows is: "My intentions somehow cause my blinkings." I hold that this is still a causal understanding, just a particularly shallow one.[77] Thus, on the account I am offering, since deviance is relative to the agent's causal understanding, any act of blinking caused in any way by this agent's intentions is successful intentional action: there is no possibility of deviant causation here. And this may seem problematic.

I want to respond to this worry by discussing a similar case of nonbasic action. Consider Mabel Pines: she is trapped in a crate.[78] Having become convinced of the power of silliness, she decides to escape from the crate by doing the silliest thing she can: jumping out of a knothole, which she believes will somehow free her from the crate.

[76] Thanks to Kieran Setiya for this problem and example.
[77] Cf. Woodward 2003 p. 18.
[78] This case is drawn from Hirsch 2012.

She only manages to get her finger out, but a passing woodpecker mistakes it for a worm and promptly pecks the crate to pieces. And lo! she is free. What I suggest is that Mabel intentionally freed herself from the crate. She believed: trying to jump through this hole will, somehow, free me from this crate. Given that she lives in the rather peculiar world of children's television, her belief was both true and justified. What this example shows is that what counts as deviance is deeply relative to the agent's thoughts about how their action is caused: almost any more particular understanding of how her action would free her would have made this a case of deviant causation. So I wish to apply the same reasoning to the case of basic actions. It is possible to merely think "My intention will somehow cause a blink," and successfully and intentionally blink.

One might therefore wonder: why don't agents always just think "My intentions will somehow cause ___," since they can thereby avoid every instance of deviance? The answer is that the aim of action is not to avoid deviance, but to be successful; having a more detailed causal understanding helps in that aim. This is just because agents' causal understanding of the connection between intention and action typically concerns the circumstances in which their intentions will be effective. Agents who lack such knowledge will be generally unable to perform successful actions, since they will try and fail far too often. As noted, the unsophisticated blinker will keep on trying to blink when his eyes are taped open. Indeed, lacking the relevant causal understanding, he won't be able to reason from "I want to blink" and "I can't blink when my eyes are taped open" to taking the tape off his eyes.

In brief: in the limiting case, while agents with absolutely minimal causal understandings can still act intentionally, in such cases we cannot distinguish between deviant and non-deviant success (though other categories, such as side effect and failure, would still apply). But such a limited understanding practically guarantees that such agents will consistently fail; almost all actually existing agents will have a far better causal understanding of the connection between intention and action.

2.3.5: A self-aware climber; Parfit on drugs

Consider a variation on Davidson's climber who is well aware of his problems with anxiety, and therefore simply plans on letting his partner fall when his hands shake. *Deviance* implies that this climber intentionally lets go of the rope. Yet that seems wrong: the climber does not drop the rope intentionally, but out of nerves.[79]

[79] Thanks to Kieran Setiya for this objection.

Or so one might argue. I respond: consider an analogous case of an agent's moving their body by nonstandard means. For instance, I might raise my right hand by picking it up with my left. If I were then to say "I'm raising my hand," my audience would think I was making a sort of a pun. And this is right: in English, a claim like "I'm raising my hand" or "I'm dropping the rope" has the connotation that I am doing so in the normal way one moves one's body. So in this unusual case, we should distinguish two questions: "Am I raising my hand in the normal way?" and "Is what I am doing intentional?" While the answer to the former is "no," the answer to the latter is (I suggest) "yes."

We can apply the same reasoning to the modified climber. While it seems strange to say "He is letting go of the rope," it seems clear enough that when the rope slips his grasp, this is intentional: and *Deviance* captures just this. The same basic idea applies to a case derived from Thomas Schelling via Derek Parfit:

> An armed robber has broken into Parfit's house, and threatens to kill his children unless he opens the safe. Parfit, thinking quickly, takes a drug which temporarily makes him completely irrational, saying things like "Go ahead. I love my children. So please kill them." The robber, realizing he has no hope of bargaining with Parfit, leaves without harming anyone.[80]

The same basic problem appears here as well. According to *Deviance*, Parfit (or at least, past-Parfit) saves the children intentionally. Yet it would seem extremely odd to say, while Parfit is raving, "Parfit is heroically saving his children." As before, I would separate out two questions: what is now-Parfit up to, and is this intentional relative to past-Parfit? Now-Parfit is raving, and being so mad may in fact intend nothing. But past-Parfit's plans are being executed to the letter; saving the children is thus intentional.

The basic idea is that there are a great many ways of going about getting what we want. And we really do want to register that something weird is going on in the weird cases. But "deviant" is not the category that we need, since in these cases, we see agents achieving their ends in exactly the way they planned to. These are intended effects: they are just odd in various other ways.

2.4: THE PRACTICAL CATEGORIES GENERALLY

Having worked through *Deviance* and its attendant complications, I want to return to the topic of the practical categories generally. I said at the outset that they represent different ways the world may depend on the agent's thought. I now want to explore the hypothesis that each category can be defined on the model of *Deviance* – that the

[80] This story is much condensed from Parfit 1986 pp. 12-13, referencing Schelling 1981.

categories can all be defined with reference to how the agent takes the world to depend on their thought. But doing this requires more specificity about the content of the agent's causal conception. I'll first discuss how to generalize from conceptions of causal dependence (my focus thus far) to conceptions of dependence generally, and then argue that *Deviance* implies a surprising conclusion: the practical thought of agents is de se, or "self-locating," in a novel sense. Given this understanding of practical thought, I'll sketch definitions of the practical categories of intended effect, side effect, and failure on that basis, then respond to two prima facie objections to that sketch.

Why worry about the possibility of non-causal dependence? As I discussed in Chapter One (§1.2.3), there is a long-running debate about whether action depends causally on intention; I don't intend to settle the question here, but will show how the account of deviance I've developed is independent of it. Consider the typical examples of non-causal actions: conventional acts like contracting, acts which are such that they cannot be done but intentionally.[81] If you don't realize that the piece of paper in front of you is a contract, then your signing it is not an act of contracting. And this throws a wrench into the idea that your intention to contract could be a cause of your contracting, since in general relations of causal and non-causal dependence are mutually exclusive.[82] But I think it is plausible that (in cases of successful contracting) your act still depends on your intention to contract; this is merely a variety of non-causal dependence.[83] Assuming that that is all correct, nevertheless the same basic features of causal understanding that I appealed to above can be imported to our account of agents' understanding of dependences: when intentional action involves non-causal dependence, the agent understands *how* the world non-causally depends on their thought. So it is relatively straightforward to generalize from causal dependence to dependence generally. If there are cases of non-causal dependence of the world on thought, then *Deviance* can be extended

[81] As discussed by Anscombe 1969 p. 61.

[82] Woodward (2014) makes this point forcefully. The basic problem is that treating necessary or conceptual connections as causal connections leads to seriously flawed reasoning. To take Woodward's example, imagine that we are investigating the influence of high-density and low-density cholesterol on heart disease. Quite obviously, we might consider a third variable: total cholesterol. But if we treat this as depending causally on high and low density cholesterol, then we are forced to condition on it when determining whether low-density cholesterol has any direct effect on heart disease. But if we do this, the necessary connections among these three variables mean that any variation in low-density cholesterol will be accompanied by variation in high-density cholesterol (since we are keeping total cholesterol fixed). This quite clearly will prevent us from accurately measuring the effects of low density cholesterol. So we cannot treat relations of non-causal dependence as if they are relations of causal dependence.

[83] While the notion of non-causal dependence was once out of favor in analytic philosophy, it has enjoyed a recent revival in the works of a number of thinkers, e.g. Kit Fine (1994), Jonathan Schaffer (2010), and Theodore Sider (2012), to name just a few.

naturally: agents must understand how the world depends, causally or otherwise, on their thought. Thus going forward I will assume that there are such cases, but nothing rides on this; the practical categories I outline are independent of this dispute.

I next want to draw out a surprising consequence of *Deviance*: agents' understanding of dependences must be de se, or self-locating, in the sense described in Chapter One (§1.2.2). That is, practical thought must be *about itself*; in particular, it must attribute causal efficacy (or non-causal grounding) to itself, e.g. "This very thought grounds A in such-and-such a way" or "A depends on *this very thought* in such-and-such a way." And such a thought would be true just in case A really did depend on that thought in that way, and false otherwise. Such thoughts would not imply that *every* such thought anywhere in the world would have such efficacy, but merely that *this particular* thought has such token-causal efficacy. And representations have such efficacy all the time – thoughts are often enough the causes of various worldly states and events. The only peculiar thing about such thought is that they also say that they have this property.

I wish to argue that the understanding of dependences involved in intentional action is de se in this particularly strong sense. The argument for this is straightforward: agents' actions depend on their understanding of dependences. As I argued in §2.2, agents' instrumental reasoning involves their understanding of how the world depends on their intentions. But we can also ask whether this connection is deviant or not. Applying the arguments of earlier sections: agents need an understanding of how the world depends on their understanding of dependences. So we can either postulate an infinite set of representations or de se representations; the latter is preferable. So agents' understandings of dependences are de se.

So, intentional action involves a de se understanding of dependences, i.e. a representational state along the lines of "This very representation grounds A in such-and-such a way," where A is anything suited to be explained by a mental state – a process, a state, an event, etc. This allows us to characterize the practical categories quite neatly.

> *Intentional*: Some worldly thing A is an intentional effect of an agent X if and only if X has a true representation of the form "This very representation grounds A."
>
> *Failure*: An agent X fails in action if and only if X has a false representation of the form "This very representation grounds A."
>
> *Side Effect*: Some worldly thing A is a side effect of an agent X's actions if and only if a) A depends on X's practical thought and b) A's existence does not depend X's recognition of that dependence.[84]

[84] Why not "...A depends on X's practical thought, but X's ends do not depend on it?" Because there may be cases where the side effects of X's actions really do contribute to X's ends, despite X not bringing them about for that reason – this property of the side effect is *practically irrelevant*, as far as X is concerned.

The basic idea in each case is straightforward. *Intentional* simply says: the agent takes their practical thought to produce some effect, and it does, and in just the way they take it to. *Failure* says that something goes wrong: their practical thought doesn't produce that effect, or produces its opposite, or produces it in an unanticipated way (this last case is, of course, *Deviance*). And *Side Effect* says that the agent's practical reasoning has produced some effect, but the production of that effect did not figure in the agent's practical reasoning.

These theses have a number of important features. First, while I have stated them as biconditionals, the arguments thus far only establish the left-to-right direction. I have not yet ruled out that there may be other, further necessary components of intentional action, such as a "distinct practical attitude,"[85] or a belief that one ought to so act, or some other feature. I therefore state them as biconditionals in anticipation of Chapter Six (§6.1), where I will attempt to establish the other direction. Second, these claims are deliberately silent about what sort of things might be substituted for A – actions, events, states, etc. They are practical, not metaphysical, categories. Similarly, these claims apply in just the same way to basic and non-basic actions. And thus *Intentional* is not itself a definition of intentional *action*, but rather explains which actions are intentional. I do however believe that an account of the metaphysics of action can be developed on this basis, and I make some tentative suggestions for how to do so in Chapter Five (§5.5). Lastly, they acknowledge Anscombe's famous claim that intentional actions are only intentional under certain descriptions.[86] In this case, the relevant descriptions are those that figure in the agent's thought about those actions.

I now want to confront two worries about *Side Effect*, one forcefully articulated by Michael Bratman, the other by Joshua Knobe.[87] On my reading, the problem Bratman raises is that the definition offered above is not even internally coherent. Why is this? Bratman's original argument was directed against simple cognitivist theories of intention, on which intentions are just a species of prediction (e.g. "I will do A"). As Bratman points out, if the agent knows "If I do A, I'll do B," then they ought to reason "So I'll do B." But if intentions are just a species of prediction, then the agent also intends to do B. But this obliterates the distinction between intended and side effects: no rational agent would ever be in the state described by *Side Effect*.

I want to respond to this worry by returning to the distinction made earlier between two kinds of de se thoughts: thoughts centered on the agent and thoughts centered on thoughts. Contrast the following two inferences:

[85] The phrase is due to Sarah Paul (2009 p. 11).
[86] Anscombe 1957 §6.
[87] Bratman 1991 p. 123. Sarah Paul presses a similar objection in her 2010 §3.

RS believes "I am RS"
RS believes "RS is P"
On that basis, RS believes "I am P"

This is, obviously, a perfectly valid inference. Now consider the same form of inference, but with representation-centered representations instead of agent-centered representations:

RS believes "This very belief is P"
RS believes "That belief is Q"
On that basis, RS believes "This very belief is Q"

This inference, on the other hand, is invalid. The reason is that the phrases "this very belief" in the premise and conclusion refer to different beliefs. The premises therefore do not even make the conclusion more probable, since they simply have nothing to do with it. Thus, the shift from agent-centered to representation-centered de se thoughts allows *Side Effect* to escape Bratman's criticism. Thoughts about side effects are de se, but centered on the agent; practical thought about intended effects is centered on itself.

Knobe raises a rather different problem. Rather than arguing that *Side Effect* is incoherent, he argues that it fails to capture the morally laden aspects of the ordinary understanding of side effect. In a now-famous experiment, Knobe presented undergraduates with two vignettes in which a CEO, focused only on profit, decides on a course of action that has the side effect (according to *Side Effect*) of (in one case) harming the environment and (in the other) helping it. When asked whether the CEO intentionally or unintentionally harmed the environment, 82% of respondents claimed it was intentional; when asked whether the helping was intentional or unintentional, only 23% claimed it was intentional; this phenomenon has become known as the side effect effect.[88] This experiment seems to rather decisively demonstrate that the ordinary conception of side effect is not morally neutral. So *Side Effect*, which is neutral, cannot capture our ordinary understanding of the concept.

I will not pretend to here decisively answer Knobe's criticism, but will instead merely suggest that the problem is not quite as devastating as it might first appear. This is because further experiments, performed by Steve Gugliemo and Bertram Malle, show that the strength of Knobe's results is the result of survey design – in particular, the use of forced-choice answers (i.e. "Was doing A intentional or unintentional?"), when a more sophisticated array of answers is appropriate to the case.[89] Gugliemo and Malle used Knobe's original vignettes the following array: "Pick the most accurate of the following: the CEO intentionally / knowingly / willingly / unintentionally did A."

[88] Knobe 2003 p. 192. The same effect has been noted by many others, including Leslie et al. 2006, Nadelhoffer 2006, and Cokely and Feltz 2009.
[89] Gugliemo and Malle 2010 p. 1642.

Given this more sophisticated array, only 1% of respondents held that the CEO intentionally harmed the environment, and 86% instead held that the CEO knowingly harmed the environment. This is because "unintentional" is not simply the negation of "intentional;" instead these two terms designate opposite ends of a spectrum of practical categories. Thus, when forced into a false dichotomy, respondents chose the "most correct" answer available. This strongly suggests that the ordinary concept of side effect is not quite as morally loaded as Knobe concluded.

However, Gugliemo and Malle nevertheless did find that the moral valence of the side effect did make some difference to the responses – e.g. respondents were in some cases more likely to say that that bad effects were intentional. I suggest that this is due to the interaction between practical categories like side effect and moral categories like blame. That is, it might very well be true that we are more blameworthy for bad side effects than we are praiseworthy for good side effects. If this is true, then respondents may be confusing the moral category of blame with the practical category of side effect. Or so I hypothesize – Gugliemo and Malle provide a related but somewhat different explanation of the data. Whichever explanation one prefers, the point is that these much smaller side effect effects are amenable to such explanations; they do not immediately demonstrate that the ordinary concept of side effect is morally laden through and through. So while there are important and unresolved questions about how practical categories and moral categories interact, I suggest that the side effect effect does not immediately refute *Side Effect*.

Thus: *Side Effect* is not internally incoherent, nor is it radically at odds with the ordinary concept of side effects. This of course does not show that it (or *Intentional* or *Failure*, for that matter) are immune to criticism. I merely suggest that my hypothesis - that the practical categories generally reflect ways the agent takes the world to depend on their thought – is not subject to immediate refutation on conceptual or empirical grounds.

2.5: CONCLUSION

This paper began with a familiar idea: as agents we impact the world in a wide variety of ways. We walk to the grocery store, get married, inadvertently reveal another's secrets, cause collateral damage in warfare, slip and fall instead of fetching the good china from the high shelf. We divide these varied impacts using practical categories: intentional, side effect, failure – and perhaps many other subspecies of the foregoing. I suggest that these categories

all reflect ways that these impacts relate to the agent's own practical thought, and in particular, how agents understand the impact of their own practical thought on the world. This is a rather strong idea, and I've hardly been able to defend it as thoroughly as it merits. Instead, I worked out in detail how it applies to one very specific (and notoriously recalcitrant) category: deviance. And while I sketched how it might apply to the other categories, and tried to defend it against two rather serious objections, I've not proven it superior to all comers, so to speak. Instead, I merely hope to have suggested the fruitfulness of this approach to the problem of practical categories. The tree from which this fruit derives is the fact that ordinary agents (I have argued) have a far more sophisticated understanding of how their practical thought affects the world than has typically been supposed. This is reflected in the fact that ordinary agents manage to navigate their worlds rather effectively; if they are doing so on the basis of means-end reasoning, then they must understanding how their practical thought affects the world.

This overall proposal suggests a striking consequence: that practical thought, i.e. intentions and means-end beliefs, represent themselves as the ground of action. These arguments uggest a novel virtue of cognitivist-style theories. Previous proponents of such theories, such as Gilbert Harman, David Velleman, and Kieran Setiya have argued that cognitivism is necessary to account for practical knowledge - the fact that, typically, if an agent is doing A intentionally, that agent knows "I am doing A."[90] Yet I have here not breathed a word of practical knowledge, but instead argued for cognitivism on the basis of its fruitfulness for understanding both practical categories and means-end reasoning. And this has typically been the great weakness of such theories: their difficulties with accounting for aspects of intentional action other than practical knowledge. I hope here to have shown how to turn that weakness into strength.

[90] Cf. Harman 1976 p. 434, Velleman 2007 p. 16, Setiya 2007a p. 26.

3.0: MEASURING THE MEANS TO OUR ENDS

3.1: INSTRUMENTAL REASONING

I'm thirsty; I want to slake my thirst. On hand, I have water, coffee, tea, and milk. But coffee would take too long, I don't like green tea, and I think the milk is just faintly sour (though I'm wrong – I have forgotten than I bought more earlier this morning). So I drink some water. This is the process of *instrumental reasoning*. The story seems to involve three key parts: an end (slake thirst), an array of means (e.g. drinking tea), and beliefs which link the two ("Drinking milk is a worse idea than drinking water"), comparing the efficacy of the various means.

Instrumental reasoning is not, at least on its face, the same as "enkratic" reasoning.[91] Consider the following story: I know I ought to visit my mother in the hospital, despite the heartache it brings. So I visit my mother. Here, we have two moving parts: a belief about what I ought to do, and the doing of it. We can see the difference between this sort of reasoning and instrumental reasoning more clearly when we consider a variant on the story: I know I ought to visit my mother, but cannot bring myself to do so. So I don't answer the phone when my sister calls, since I know if I hear her voice I won't be able to say no. Here, we have instrumental reasoning in the service of an end I know to be wrong. The end is avoiding my mother; the means is avoiding my sister; the instrumental belief concerns what would happen if I answered the phone. The fact that these sorts of reasoning can come apart, operate at cross purposes, as well as the fact that their internal structures differ both suggest that enkratic reasoning is distinct from instrumental reasoning

It is instrumental reasoning, rather than enkratic reasoning, that is our topic here. As agents, we tend to get around in the world pretty well without thinking too hard about this process; as philosophers, we want to investigate its properties more reflectively. For instance: what sorts of combinations of its three parts are *possible*? Is it even possible to intend some end E, believe that M is a necessary means to E, and not intend M? Or does the fact that I

[91] The term but not the idea is due to to Broome (2013).

am not taking the necessary means demonstrate that I don't really intend the end?[92] Given that it is possible, is it *rational* to be in such a state?[93] If not, how should one get out of it – what prescriptions does instrumental reason deliver? Or are such prescriptions mythical; is there no such thing as instrumental rationality?[94] These questions concern what I will call *instrumental principles*: principles which tell us something about schematic combinations of ends, means, and instrumental beliefs.

One way of investigating such principles is to begin with the three sorts of objects they relate: ends, means, and instrumental beliefs. If it is irrational to intend the worst means to one's ends, and rational to intend the best, that must be because of some features of these combinations of states.[95] For example, consider an argument extracted from the work of R. Jay Wallace and Kieran Setiya, which focuses on a specific component of instrumental reasoning: beliefs that M is a necessary means to doing E.[96] Now of course one may have such beliefs about other people ("John needs to get a hammer if he wants to finish that birdhouse"), or about one's future self ("I'll need to pack at least three liters of water if I want to make it through this hike"). But we can also have first-person, present-tense such beliefs: "I need to leave *right now* to catch that bus." And such beliefs are implicitly about intentions, not actions. Consider: it's true that I need to keep on breathing in order to stay alive. But fortunately I don't need to intend to breathe in order to breathe; mostly I just do it automatically. My beliefs about necessary means concern those actions that I must intend to do if I am to do them.[97] Taken together, these points imply that beliefs about necessary means have the following form: "If I succeed at E, then I intend M now." Given this conception of instrumental beliefs, we can make the following argument for an instrumental principle:

> *Wallace-Setiya*: Given that I intend E, then I ought to believe "I might succeed at E."[98] Given that and a belief that M is a necessary means to E, I ought to believe "I might now intend M." But I should only believe that if I intend M.[99] So it is irrational to intend E, believe M is a necessary means to E, and not intend M.

Thus this argument begins by investigating the nature of instrumental beliefs, and from the content of such beliefs (plus some other premises) derives an instrumental principle.

---

[92] Finlay 2010.
[93] Cf. the debate between Broome and Kolodny (Broome 1999, Kolodny 2005, Broome 2007, Kolodny 2007).
[94] As e.g. Raz (2005) argues.
[95] Cf. Schroeder 2004 p. 341 and Setiya 2007b p. 655.
[96] Wallace 2001 p. 1, Setiya 2007b p. 651.
[97] Wallace 2001 p.21; Setiya 2007b pp. 667-668.
[98] Wallace 2001 p. 20; Setiya 2007b p. 664.
[99] Wallace 2001 p. 21; Setiya 2007b pp. 670-671.

One might object to this argument for a number of reasons. It relies on quite controversial claims about self-knowledge and about rationality. But even setting those problems aside, the argument is inherently limited, insofar as it relies on the notion of *necessary* means. Typically instrumental reasoning does not involve such beliefs; as in the story that opened this chapter, we are often confronted with a wide array of potential means none of which are necessary for our ends. One natural suggestion is to introduce a notion of sufficient means, but unfortunately this will also not capture the full range of instrumental reasoning. Often enough we lack any strictly sufficient means to our ends: I want to please my guests at dinner tonight. But I'm not sure what they like to eat, so I take my best guess. Do I believe that meatloaf is sufficient to please my guests? No; it just seems like the best bet on this wintry evening. This suggests that we need to expand our account of instrumental beliefs to account for thoughts like "Doing M is a good way of doing E" and "Doing M is a better way of doing E than M*".[100] Unfortunately, many discussions of instrumental principles restrict themselves to the simpler case of necessary means. But it remains an open question whether any such account can be generalized to the full domain of instrumental reasoning. Does the Wallace-Setiya argument, or anything like it, remain valid when we transition from believing that M is a necessary means to E to believing that M is the best means to E?

On my view, the answer is yes. But demonstrating this requires a careful investigation of the content of instrumental beliefs generally. Thus I turn to that research program which has investigated the details of instrumental reasoning most carefully – decision theory – and try to extract some answers from the thickets of formalism found there. In what follows, I will discuss and criticize evidential decision theory, causal decision theory, and their ratificationist variants, and argue that they are ultimately unsatisfactory as they stand: they cannot account for instrumental reasoning in cases where success in action depends on the agent's reasons for action. Instead, instrumental beliefs must be about the very situation that acting on them would involve; this implies that they are self-referential: they represent themselves as the conditional causes of actions. Given this content, an argument analogous to the Wallace-Setiya argument can be constructed for the case of instrumental reasoning generally.

[100] We might also investigate into another mostly overlooked belief: "I can do A." I'll gesture at how my account incorporates such beliefs in §3.5.3.

3.2: EVIDENTIAL DECISION THEORY

But that conclusion is a ways off yet. For now, let us consider evidential decision theory. This theory says that the rational agent takes the means with the highest utility, where the utility of a means is defined by this equation:[101]

$$v(M) = \sum_S p(S / M)u(M \& S)$$

Here, M is some means you could take – and the overall set of such means, **M**, is exclusive and exhaustive: you are guaranteed to do one and only one. S is an overall state of the world; again, **S** is exclusive and exhaustive. And what the equation says is: for a given act M and state S, first determine what your subjective expectation in S is given that you do M, then multiply that by how desirable it would be for S to be true while you do M. Then repeat this for every state, and add these results up. Repeat this process for every potential act. Then do the M with the highest expected utility.

One might object to this theory for a number of reasons. For one, it seems too cognitively demanding. It seems to imply that every time an agent takes an action, they are performing the calculation just described for every possible means and every end. And it is not plausible that this is going on every single time any agent takes an action. Some decision theorists have responded to this problem by holding that a good reasoner only reasons *as if* they were performing some such calculation; there's no need for them to actually do it as described.[102] But at least for my purposes, this is an unsatisfactory response. The hope was to offer some illumination not merely of ideally rational agents, but to understand how everyday agents think about the means to their ends.

Thankfully, I think there are ways of avoiding this problem that don't involve collapsing into pure nominalism. The initial description of EDT can be modified: rather than always thinking about all possible ends and means, we can allow that in given stretches of instrumental reasoning, the agent only considers some of the available ends and means. This makes the procedure significantly more cognitively tractable; for instance, in the initial example, the agent only compares four means and one end. In this more restricted case, the above equation boils down to the following thought: for each means, consider what your expectation in "I achieve E" is given that you do M. Then do the M that most strongly indicates E – that, in essence, would be the strongest evidence that "I achieve E" is true.

[101] I take this formulation from Joyce (1999 p. 119); the theory is due to Jeffrey (1983).
[102] Cf. Joyce 1999 Chapter 2 §6 for further discussion of this problem.

This story might sound odd: why are we talking about the *evidence* M would provide for E? But things are more straightforward than they seem. Consider the initial story: there is some agent X with the end of slaking thirst and a variety of possible means available. We can ask ourselves: if we discovered that X was sitting still and doing nothing, how much evidence would that provide for us that X's thirst was going to get slaked? Answer: not very much. By contrast, the discovery that X was drinking water would provide strong evidence for thirst-slaking. So according to evidential decision theory, drinking water is a more effective means of slaking X's thirst than sitting still. And this same basic procedure can be applied first-personally: I can ask "How confident would I be in 'I will slake my thirst' on the assumption that I drink water?" And the means that would provide the most evidence for success is, by my lights, the best available means.

There is however a problem with this way of putting things. Imagine that I am playing pool, and am considering whether to a) make an easy shot or b) make a much more difficult shot which would win the game immediately. As stated, the above account says: first, determine how much evidence making the easy shot would provide for winning the game (some). Now determine how much evidence making the difficult shot would provide (a lot, since it would win the game immediately). So the rational agent would take the more difficult shot. But this is the wrong verdict, at least on the assumption that I am almost certain I will miss the difficult shot if I try. Making the difficult shot is not something I can just do straightaway. In order to account for these sorts of problems, EDT requires that the acts under consideration be completely under the agent's control. However it turns out to be somewhat difficult to say just what such acts are – are they basic actions? Tryings? Intentions?[103] Call this "the problem of acts" -– the problem of finding some class of actions that is always completely under the agent's control to be the subject matter of instrumental beliefs. I'll argue in §5.3 that the problem is insoluble so long as we remain within the confines of standard decision theory. But let us set it aside for the moment, and grant the decision theorist their acts – I'll construe them as tryings – to see what else we can learn from decision theory.

[103] A basic action is some action which is not done by doing any other intentional action. Basic actions are thus often thought to be bodily movements, like raising one's arm. But in many case (consider for instance muscle fatigue) even the movements of one's body may not be entirely under one's control. Depending on how tryings are construed, a similar problem may arise. And I will later present a case where even the agent's intentions are not completely under their control.

Given these points, we can be a bit more precise about what account of instrumental beliefs we can extract from EDT:

> EDT: The content of "M is the best means to E" is "My credence in "I will do E" conditional on "I am trying to do M right now" is higher than it would be conditional on "I am trying to do M* right now" for any other M*."

This raises a further objection. On its face "My credence in "I will do E" conditional on "I am trying to do M right now" is higher than it would be conditional on "I am trying to do M* right now" for any other M*" is simply not synonymous with "M is the best means to E."[104] So how can it be an analysis of the content of instrumental beliefs? I suggest that we should distinguish between a linguistic investigation into how we communicate information about means-end relationships in English (or whatever other language) and an investigation into how information about means-end relationships is represented in the minds of agents. Sentences like the one above belong to the second investigation, not the first; they are not meant as a replacement for English sentences like "This is a better means than that;" such sentences function pretty well as-is, and don't need replacing. Rather, the above sentence is meant to characterize the representational content of thoughts about means-end relationships. The point here is to do so in a precise way, so as to capture their inferential connections and truth-conditions exactly. And we should not expect such connections and conditions to be simply and straightforwardly expressed in English – just imagine trying to communicate the full representational content of your current visual experience in English. So this kind of complexity is no objection to a theory of the content of instrumental beliefs.

Similarly, EDT does not imply that agents are always consciously or explicitly having such thoughts when they act. Thoughts about good means are rarely occurrent thoughts – when I decide what to drink, I don't need to consciously run through each possible choice. Instead, these are background beliefs that structure my actions and instrumental reasoning. And EDT does a surprisingly good job of accounting for the nature of such beliefs, at least in a wide range of cases. But it also breaks down in some crucial ways, as I will now show.

[104] Thanks to an anonymous reviewer for making this objection.

3.3: CAUSAL DECISION THEORY

There is a well-known reason why EDT, at least as thus far developed,[105] is not a satisfactory account of instrumental beliefs. It goes wrong in cases where correlation does not imply causation, e.g. where there is some *common cause* of tryings and good outcomes; one well-known example is known as the "smoking lesion."[106] Consider: in our world, smoking and lung cancer are highly correlated, because the former causes the latter. But in a somewhat different world, this correlation is instead explained by the fact that they are both frequently caused by an underlying genetic factor; there is no direct *causal* connection between them. Now imagine a man in that world deciding whether to smoke. He knows these facts about his world, and so a decision to smoke is strong evidence that he has the gene, and therefore that he will get cancer. According to EDT, the man should not smoke, since that is evidence that he will get cancer. But this seems wrong: whether or not he chooses to smoke now has no possible impact on whether he has the gene, and therefore no possible impact on whether he gets cancer. So he should enjoy a cigarette.

Causal decision theorists respond to this problem by distinguishing between backtracking and nonbacktracking reasoning.[107] In the above case, when the agent considers what would happen if they intended to smoke, they reason backwards from intending to smoke to its genetic cause, and then forwards from that gene to cancer. But this is a bad way to determine the causal effects of smoking. To do that we must use nonbacktracking reasoning instead: we must suppose that smoking occurs independently of its normal causes. In the case above, under the supposition that he smokes, the man should reason forwards to that action's effects (the pleasures of smoking), but not backwards to its genetic cause. So the man can correctly reason that smoking is a good means to pleasure. Generalizing these ideas, we have:

---

[105] A number of decision theorists have responded to the following problem with variations on a doctrine known as "ratificationism" (e.g Eells 1982, Jeffrey 1983, and Price 1986 and 1993); I will return to this topic after discussing causal decision theory.

[106] This objection derives from William Newcomb via Robert Nozick (1969). It is now widely but not universally accepted among decision theorists as a refutation of unvarnished evidential decision theory (cf. Weirich 2012 §4). The example that follows is a "medical" Newcomb problem.

[107] This terminology is due to Lewis (1979b p. 456), and characterizes causal vs. non-causal supposition generally, not just in the case of instrumental reasoning.

> CDT: The content of "M is the best means to E" is "Were I to try to do M right now, that would be more likely to cause E than would any other M*."

As James Joyce puts it, "[this] gauges the extent to which performing [M] can be expected to *bring about* desirable or undesirable outcomes."[108]

But, I suggest, CDT is also unsatisfactory. It fails to account for straightforward reasoning in cases where success in action depends on the agent's reasons for action. Suppose I wish to roll a rather large boulder down a hill. Since I'm not strong enough to budge it myself, I need a friend to help me. But my friend (who cares a great deal about my reasons for doing things) will only help me push the boulder if I do so in order to roll it down the hill. So: is pushing the boulder a good means to rolling the boulder down the hill? According to *CDT*, no. When I am evaluating the efficacy of trying to push the boulder, *CDT* says: imagine that your trying to push the boulder were set by intervention, and see how likely the boulder would be to roll down the hill. But when I suppose that my trying is *set by intervention*, and *not* by my own reasoning, I therefore suppose that I am not pushing the boulder in order to roll it – I'm just pushing it for the heck of it. So my friend won't help me and the boulder won't budge. But this, I take it, is the clearly wrong answer. Pushing the boulder is a great means to rolling it, since my friend will come help me and together we'll move it.[109]

How might a causal decision theorist try to respond to this objection? One answer is to turn to ratificationism, but first, let's consider some other possible responses. First, one might argue that the claim "I am doing A on the basis of B" is just one of the states of the world the agent has credences in – that the agent takes account of this fact in the same way they take account of any other fact about the world (e.g. that boulders roll down hills). So I should take as given "I am pushing the boulder in order to roll it," and since given that pushing the boulder will be effective, decide to push the boulder. The problem of course is that such reasoning is quite backwards. When I am considering whether to push the boulder, I am plainly not already pushing the boulder in order to roll it – and I know this. So on this account, I should expect the boulder not to move. So we are still left with the mistaken verdict that pushing the boulder is a bad means to rolling it.

So perhaps instead we should consider the truth of "I am doing A in order to do B" to be a *consequence* of intending A. I know that if I were to try to push the boulder, then it would follow that I was pushing it in order to

---

[108] Joyce 1999 p. 161.

[109] One way to read this objection is as revising a famously obscure argument due to David Hume known as "Hume's Circle" (1740 §3.2.1.17); see Anscombe 1969 and 1978 for some discussion. But discussing the relations between these arguments (and how the account of instrumental beliefs developed here offers a way out of the Circle) is beyond the scope of this chapter.

roll it. So trying to push the boulder would be efficacious. Unfortunately, this reasoning is also problematic: it is simply implausible that intending A on the basis of B is a consequence of intending A. After all, as noted above, according to CDT, I am already supposing that my trying to push the boulder has no connection with my end of rolling it down the hill. It follows from this that I am not pushing the boulder in order to roll it. So we are still left with the mistaken verdict.

So perhaps instead the causal decision theorist should hold that "I am doing A in order to do B" and "I am doing A for no particular reason" are simply different acts, and should be treated separately in the theory.[110] Or, in other words, in order to capture the reasoning in the boulder-rolling case, I should consider the effects of an intervention on "I am pushing the boulder in order to roll it." And then we get the result that the boulder will probably roll, and so this is a good means to rolling the boulder.

One problem with this response is that, in my view, "I am doing A on the basis of B" is in part a causal claim – it asserts at least that my goal of doing B is a cause of my doing A. But, at least in my view, there is no well-defined concept of causal supposition of causal claims – "If it were the case that A causes B, then..." is not a well-defined proposition.[111] But set my metaphysical quibbles aside; let us assume that these claims make sense. If that's so, then the best way to make sense of them is some variant of Lewis' nearest-possible-world analysis.[112] That is, in order to determine whether "If it were the case that p, then q" is true, we look to the world that most resembles the actual world in which p true, and see whether q is. In the boulder-rolling case, I would reason as follows: "In the nearest world where I push the boulder in order to roll it, my friend comes and helps me. So pushing the boulder is a good means to rolling it." So this proposal can account for the boulder case.

But let us consider how this would apply in a more complicated scenario. Imagine that Marina has lent me her pen, and I have promised to give the pen she lent me back later that afternoon. Now it is later that afternoon, and I am wondering "Is giving Marina this pen a good way of keeping my promise?" Since it is the pen she has lent me, the answer is clearly "yes." How would this proposal handle this case? We would perform some intervention on "I am giving Marina this pen in order to keep my promise," and see whether that is a way of keeping my promise. But unfortunately it is not – we have now separated out that piece of practical reasoning from its typical cause, *the act of*

[110] Thanks to James Shaw for suggesting this style of response.
[111] In essence, this is because I follow James Woodward in holding that causal claims and causal suppositions are defined by the notion of intervention, and there is no coherent notion of intervening on a causal invariance as opposed to the value of a particular variable (Woodward 2003 Chapter 3).
[112] Lewis 1979b pp. 464-465.

promising. The counterfactual situation we are considering is one in which my intention to return the pen has nothing to do with my actually promising to give it back. And this prevents it from being an act of keeping my promise. Keeping a promise is more than just the conjunction of promising to do A and doing A – if for instance I've completely forgotten about my promise, and do A for some other reason, I haven't fulfilled my promise. So this proposal wrongly entails that giving Marina this pen is not a way of keeping my promise.

Perhaps the causal decision theorist would response by further expanding the scope of the causal supposition – I need to consider whether "If I were to give Marina this pen on the basis of my belief that this is a way of keeping my promise, and if I were to believe that on the basis of having promised to give Marina her pen, then I would keep my promise to Marina" is true. And it probably is. So this proposal can get the right result about giving Marina her pen back.

But now consider what happens if I wonder whether giving Marina some other pen is a way of keeping my promise. We need to look to the nearest possible world in which "I give Marina pen* on the basis of my belief that giving her pen* is a way of keeping my promise and I believe that on the basis of promising to give Marina her pen" is true and see whether "I fulfill my promise" is also true. And probably the smallest change we need to make in order to make the first claim true is to let pen* be the pen Marina gave me in the first place.[113] And this gives the result that giving Marina pen* in the actual world is also a good way of keeping my promise, which it plainly isn't, since it is not the pen she lent me. So this proposal also fails – it will wrongly count far too many things as good means to my ends.

So I conclude: there simply is no way for unvarnished causal decision theory to handle cases of instrumental reasoning where success in action depends on my reasons for action. I should emphasize that the cases I've considered are not meant to be particularly weird, the way that typical objections to CDT are; rather, the reasoning in these cases is meant to be as straightforward as it comes. Pushing the boulder is plainly a good way of rolling it; giving Marina her pen back is plainly a good way of keeping my promise. The convolutions I've canvassed are not inherent in the cases, but lie merely in the attempts of CDT to capture such reasoning with its limited resources.

[113] We can, if we like, have Marina's act of giving me a pen be an effect of my promising to return whatever pen she gives me (she won't hand it to me until she knows I'll give it back). This ensures that the choice of pen is not temporally prior to the things we are supposing changed in the alternate possible world, and thus that my argument does not depend on illicit backtracking reasoning.

3.4: RATIFICATIONISM

Many a decision theorist might suspect that I have been saving the best, most comprehensive decision theory for last: ratificationism.[114] Ratificationism was originally developed by evidential decision theorists in response to the sort of problem I mentioned earlier. Unfortunately, it doesn't quite manage to solve all the relevant problems.[115] But has also been adopted by causal decision theorists to handle cases where a decision to do M provides evidence about the efficacy of doing M. Perhaps the original such case is known as "Death in Damascus," and it runs as follows: one morning, Death comes to a resident of Damascus, and tells her that (like it or not), she is now playing the following game: Death will kill her tonight, unless she manages to avoid him.[116] He has already guessed how she will try to avoid him, and having played this game innumerable times, he is almost invariably correct. So later tonight he will go to the place he has guessed she will be, and if she is there, kill her. He leaves, and she has three options: go to Samara, stay in Damascus, or pick a destination by flipping a coin. Let's say that, prior to deciding, she believes that Death has predicted she will stay in Damascus (so as to be with her family). So she decides to go to Samara. But *this decision* is evidence that Death has actually predicted she will go to Samara, so she ought to stay in Damascus instead. But *that decision* is evidence that Death thinks she will be in Damascus, so... – and so on. What this shows is that neither going to Samara not staying in Damascus is *ratifiable*: choosing it provides evidence that the other choice would be more effective. Her only ratifiable option her is to flip a coin and hope for the best.

James Joyce puts the idea more formally: an act M is ratifiable only if the agent "regards [M] and [the decision to M] as better news than [M*] and [the decision to M*] for every other act [M*] under consideration."[117] As with EDT, the basic idea can perhaps be illustrated more clearly by thinking about how we would apply it to some third party. We wonder: how likely would it be for M to cause E, on the supposition that the agent decides to do M and does M. And we compare this with how likely M* would be to cause E, on the supposition that the agent decides to do M but does M* instead. To apply this to the Damascus case, if we discovered that the agent had decided to go to Damascus, that would be evidence that Death would be in Damascus, and thus that it would be best

[114] Though recently ratification has been subject to objections based on somewhat tricky counterexamples; I'll discuss one such case later in §3.5.2
[115] See Joyce 1999 Chapter 5 §2 for some discussion.
[116] Harper 1986 §2.
[117] Joyce 1999 p. 158. I've changed Joyce's notation slightly to bring it in line with my own usage. I should also note that ratification has the odd consequence of introducing decisions to act into the theory, but as with the problem of acts, let us ignore this for the moment.

if the Samaritan nevertheless stayed in Samara. However, if we discovered that she had opted to flip a coin, that would provide no evidence about where Death was, and so a decision to go to either Samara or Damascus would not appear to be better news than a decision to abide by the coin flip. So only flipping is ratifiable.

Now, decision theorists usually only have recourse to ratification when there are unratifiable options. But we can instead think of ratification as providing a general account of how agents think about the efficacy of means, as follows:

> *Ratification*: The content of "M is the best means to E" is "My credence in "Trying to do M will cause E" conditional on a decision to do M is higher than my credence in "Trying to do M* will cause E" conditional on a decision to M*, for all M*."

This suggests a possible solution to the cases I deployed against CDT. Suppose that I decide to push the boulder. This, as I argued, can hardly be a *cause* of "I am pushing the boulder in order to roll it," but it is *evidence* for that proposition. So, on the supposition that I decide to push the boulder, pushing it will probably cause it to roll (since my friend will probably help me).

But I do not think this is a general solution. Often enough, agents have multiple potential reasons for action. I might give Marina her pen on a whim, or in order to keep a promise, or in order to make myself look good. So the discovery that I decided to give Marina her pen would, at best, be equivocal evidence for all of these propositions. So I should – on this account – be quite uncertain about whether giving Marina her pen back is a good means to keeping my promise. But that's simply false: giving Marina her pen is an *excellent* means to keeping my promise, even if I have a variety of other possible motives.[118]

I conclude that ratification is also not a workable solution to the problem of the dependence of success on reasons. So what is?

3.5: SELF-REFERENCE

I've argued that extant decision theories fail in cases where success in action depends on the agent's reasons for action. Now it's time to say how I think such reasoning ought to be accounted for. I'll first argue that instrumental beliefs represent themselves as the causes of actions, conditional on those actions being instrumentally efficient.

[118] One might be tempted to try a move from before – to insist that the agent asks "How like is A to cause B on the supposition that I decide to do A on the basis of B?" But this will run into just the same problems as before.

And I'll then show how this properly accounts for instrumental reasoning in a variety of problem cases. Moreover, this account has two other virtues: first, it suggests a solution to the problem of acts, and second, it shows how the Wallace-Setiya argument might be generalized.

3.5.1: The details

Here is one way of understanding what is happening in such cases: the nonbacktracking reasoning appealed to by CDT asks the agent how likely E would be to occur if a trying to M were *set by intervention*: as if God came down and altered the causal order such that the trying to M occurred independently of its normal causes.[119] This feature of causal supposition is essential to prevent the backtracking reasoning that sunk EDT. But this is exactly the problem: the agent *knows already* that their intentions are not going to be set by divine intervention: they are going to be caused by the agent's *own practical reasoning*, i.e. by their own estimation of how good a means to E M is. While the details of ratificationism differ, the same basic problem remains: a mismatch between the situation that instrumental beliefs are *about* and the situation that *actually acting on them* would involve. And, I suggest, the way out of this mismatch is simple: the causal hypothesis that each instrumental belief must entertain is "How likely is E to come about if *this belief* causes me to do M?"

Yet one might wonder: is self-reference really necessary to account for this reasoning? Couldn't we instead use the following model: the agent has some estimate of the efficacy of M (call it IB_1), which accords with *CDT*. And then they have some further estimate of the efficacy of doing M on the basis of IB_1 (call this IB_2), which allows the agent to reason correctly in cases like the boulder-rolling case. The problem here is that this model invites a vicious regress: if the agent actually does M on the basis of IB_2, not IB_1, then IB_2 is actually the wrong estimate to act on, since it is about the efficacy of acting on IB_1. What we *really* want is an IB_3 which tells the agent the likelihood of success if the agent does M on the basis of IB_2. And now we clearly have a regress going; the only way to cut it off is to admit that instrumental beliefs are about how likely success is if the agent acts on that very belief. *Any other prior hypothesis will simply fail to be about the posterior situation.*

I hold that instrumental beliefs have as their content "This very belief will cause E by causing M, conditional on that being instrumentally efficient" (and of course I may have a higher or lower credence in this

[119] Lewis himself characterizes nonbacktracking supposition in terms of "small miracles" (1979b p. 468). But the divine terminology is inessential, and merely highlights the very generic features of nonbacktracking counterfactuals that are common even to accounts that thoroughly reject Lewis's final analysis of such counterfactuals in terms of similarities between possible worlds (e.g. Woodward 2003 §6).

belief). But one might think that the arguments thus far at best establish that their content is "*If* I do M on the basis of this very belief, then that will cause E." But despite being self-referential, such beliefs cannot properly account for reasoning in the cases I pressed against decision theory. Consider: what conditional is being used in that claim? It cannot be the material conditional; that would allow me to have *any* beliefs about those means I don't take. Perhaps therefore it is a subjunctive conditional: "If I were to do M on the basis of this very belief, that would cause E." But now we are stuck with the problem of accounting for the promising case I discussed earlier: the problem that the conditions in the supposed world may relevantly diverge from the actual world. The simplest way to avoid any such divergence is to make instrumental beliefs simply be about the actual world: such beliefs are about their own actual causal powers. They assert that they will cause action, on certain conditions.

Why only on some condition? Because we want agents to be able to have a consistent array of instrumental beliefs in ordinary situations. Consider that agents almost always have multiple means available to them, and have instrumental beliefs not only about the option they take, but also the many options they don't take. If instrumental beliefs were unconditional, they would be thoroughly inconsistent in such situations: so we must find some conditions. Given that instrumental beliefs assert that they will cause one of these options, we want to find conditions that allow agents to be consistent: only one of the means should meet the condition.[120] And we want a condition that allows agents to be rational – if we pick a condition that sometimes does not apply to the best means, then actually doing the best means would be inconsistent with the agent's instrumental beliefs. So the right condition is: "I expect this to be the most efficient means," where an efficient means just is the means most likely to cause this end without decreasing my expected success at other ends. Thus we have:

> *Self-Reference*: The content of instrumental beliefs is "This very belief will cause E by causing M, conditional on my expecting this to be the most efficient means."

And my expectation that M will be efficient just is my credence in that belief.[121]

[120] This is actually not quite right: consider the situation of Buridan's Ass: agents may sometimes take multiple means to be equally good. Addressing this problem would complicate the account here given in ways mostly irrelevant to what follows, so I will – in the main text – set it aside. But the basic procedure would be to reframe instrumental beliefs: instead of claiming to cause the best option, they would claim to rule out the bad ones. In Buridan cases everything but the equally best means would be ruled out.

[121] It's important to note that agents also have instrumental beliefs about doing nothing. For instance, perhaps at this point I the best option is just waiting around.

3.5.2: The problem cases

I should now say exactly how this account of instrumental beliefs correctly captures the instrumental reasoning in all three of the problem cases I've discussed – the smoking lesion, the rolling boulder, and the appointment in Samara.

Take the smoking lesion first. The basic idea here is that there is some common cause of smoking and lung cancer. Thus smoking is evidence that one will get cancer, despite having no tendency to actually cause cancer. In such a situation, one ought to smoke if one desires it, since that has no chance of causing cancer. The self-referential account of instrumental beliefs handles this in a straightforward way, since – like causal decision theory – it relies on the notion of causation to define the effects of various acts. None of the agent's instrumental beliefs should attribute to themselves the property of causing cancer, since the agent knows them to be causally irrelevant to that outcome.

Consider next the boulder-rolling case. Here, I am considering whether to push a boulder in order to roll it down a hill. Alone, I can't budge it, but my friend will help me on the condition that I push the boulder in order to roll it. How does this account of instrumental beliefs yield the conclusion that yes, pushing the boulder is an excellent means to rolling it? That is: why is it rational to assign a high credence to "This very belief will cause me to roll the boulder by pushing it, conditional on that being the best means?" Well, assume that that condition is met. So the belief will cause me to push because I think that's the best way to roll it. So I'd be pushing it in order to roll it. So my friend would help me, and the boulder would roll. So I should have a pretty high credence in that belief – and so pushing the boulder is a good means to rolling it. So this account of instrumental beliefs can make sense of how we reason in cases where success in action depends on reasons for action, including the analogous case of promise-keeping.

Lastly, consider the appointment in Samara. In this case, I am faced with a choice between staying in Damascus, going to Samara, or flipping a coin to decide where to go. I know that Death has almost certainly predicted which choice I will make, and has acted accordingly. Here, we want the result that the first two options are bad means, and that the third is, well, a coin toss. So how much credence should I have in "This very belief will cause me to avoid Death by going Samara, conditional on this being the best means?" Well, assume that the condition is met, and so the belief causes me to go to Samara. But if that is the case, the Death has anticipated my actions by going to Samara. So I won't avoid death. So I should assign a pretty low credence to that belief; going to Samara is a bad way to avoid death. By contrast, if I flip a coin, I'll choose randomly (as will death), so I'll have a

fifty-fifty chance of meeting Death. So I should assign a credence around .5 to "This very belief will cause me to avoid Death by flipping a coin, conditional on this being the best means." So flipping a coin turns out to be the best means – which was the desired result. So this account of instrumental beliefs can capture exactly the sort of reasoning that ratification was introduced to capture.

One might also wonder how this theory applies to cases like the Murder Lesion – a case designed so that *no* option is ratifiable. The situation in Murder Lesion is as follows: you live in a country a) ruled by a tyrant and b) where approximately one quarter of the population has a brain lesion that makes them i) willing to kill, unlike anyone else and ii) totally incompetent at actually killing, unlike anyone else. You are considering killing the tyrant; if you kill him, life will improve significantly in your country; if you try and fail, things will get worse. This situation is somewhat unfortunate. If you decide not to kill him, that is strong evidence that you don't have the lesion, and thus evidence that you would kill him if you tried, thereby making life much better. But if you decide to kill him, that is strong evidence that you *do* have the lesion, and thus evidence that you won't kill him if you try, and thus evidence that you shouldn't try to kill him. Thus, neither of your options is ratifiable. You've got no coin available, and you need to choose whether to pull the trigger *right now*. What should you do?

Plainly enough you are in a bit of a bind. And frankly, the account of instrumental beliefs developed here won't get you out of it: *you really are in a bind.*[122] The problem here is not an artifact of some particular theory of instrumental beliefs, but inherent in the situation.

3.5.3: The problem of acts

I earlier discussed "the problem of acts" – the problem of finding something completely under the agent's control (perhaps a basic action, or a trying, or an intention) to be the subject matter of instrumental beliefs. On my view, the problem as stated is insoluble. On the one hand, traditional theorists have wanted acts to be conceptually separate from instrumental beliefs, so that they could be the subject matter of such beliefs. On the other hand, it must be guaranteed that the agent performs the act if they think it best to do so. But if acts and instrumental beliefs are conceptually separate, then any connection between them must be a causal connection. But causal connections are by their nature contingent. This tension has led decision theorists to push acts further and further back into the mind

[122] Cf. Joyce's discussion of this case: he argues, quite carefully, that the agent in such a case has got inconsistent beliefs (2012 pp. 131-132).

– Joyce, for example, defines them not as acts at all, but as all-things-considered desires.[123] But no matter how far we retreat into the mind, we will never overcome the basic incompatibility between the two noted requirements.

The account of instrumental beliefs I here developed avoids this problem. An instrumental belief says "This very belief will cause B by causing A, conditional on this being instrumentally efficient." If I am uncertain about the connection between A and B, that will plain enough get built into my credence. But any uncertainty about whether *that belief* will cause A in the first place would also get built into my credence. In this way, instrumental beliefs can represent *any* uncertainty about whether I will perform *any given action* without recourse to some class of actions which are completely under my control. Rather than being about the causal powers of some separate thing which is yet somehow guaranteed to be available, they are about *their own* efficacy. And thus they are always guaranteed to have something to be about – namely themselves.

This allows for a significantly more straightforward theory. Rather than positing some special class of actions to be the objects of instrumental beliefs – a class of actions that becomes more mystical and mysterious the longer one contemplates them – instrumental beliefs can simply concern everyday actions, like "walking to the store" or "sinking the eight ball," even if the performance of those actions is uncertain and risky. Any uncertainty about performance is already contained in the instrumental belief.

It also allows a nice explanation of how agents take account of their own frailty. Consider a situation in which I am uncertain whether I could bring myself to do A even on the condition that A is the best available means. Perhaps we are locked in a struggle with a Great Old One, and one potential means to saving the world is to gaze directly into the eye of Azathoth and smite it with the Golden Sword of Y'ha-Talla. Unfortunately, I suspect that even if that were the best available choice, I would simply fail to even try, since Azathoth's gaze pierces even the mind and saps the will of all who would strike it. On the account of instrumental beliefs I've offered, this should leave me with an extremely low credence in "This very belief will cause me to smite Azathoth on the condition that it is instrumentally efficient," which just means that smiting Azathoth is in fact a terrible means, and maybe I should consider some other options that don't involve degenerating into a gibbering mess as the abyss gazes back into me.[124]

---

[123] Joyce 1999 p. 22.

[124] The account of instrumental beliefs developed here also provides a nice conception of what it is to believe "I can do A." But detailed discussion of this topic is beyond the scope of this chapter.

3.5.4: The Wallace-Setiya argument

Earlier I discussed an account of instrumental rationality due to Wallace and Setiya. While their argument was restricted to the case of necessary means, the account of instrumental beliefs developed here shows (with one important caveat) how it might be extended to the general case. Here is how this works:

Consider an agent who intends E and believes that M is the best means to E. Given the account of instrumental beliefs developed here, this means that they have a belief which represents itself as the cause of M, conditional on M being the best means. Since they believe M is the best means, they should believe that they intend M. But they should only believe that if they intend M. So an agent who intends E, believes M is the best means to E, and does not intend M is irrational.

Now, as before, this argument relies on both a controversial principle connecting intentions and beliefs, and controversial principles about rational coherence. But given those principles, it does suggest how the Wallace-Setiya argument is not restricted to the case of necessary means. There is however one sticky wicket here: and that is the possibility that the agent foresees their own frailty – they suspect that the connection between their instrumental beliefs and their intentions to be significantly less than certain. Consider again the above example involving Azathoth. Let's assume that, despite the dangers inherent in looking into Azathoth's eye, it remains the best available option (actually striking the eye is certain to save the world, at least for now). But I am still extremely uncertain that my practical reasoning will be efficacious – perhaps I estimate the likelihood of intending to strike Azathoth conditional on its being the best idea at 1%. If this is so, then I should not believe full-out that I intend to strike Azathoth, but merely have a .01 credence in that proposition. And now the next step of the Wallace-Setiya argument seems more difficult: should I only have such a credence if I actually intend to strike Azathoth?

Answering this question is far from straightforward, since it's not clear how norms of self-knowledge apply in the framework of partial beliefs. But then, giving an account of such norms has always been a necessary part of the overall Wallace-Setiya argument. This just emphasizes that if we want to extend that argument to instrumental reasoning in general, we also need a more general account of self-knowledge.

3.6: CONCLUSIONS

We can, if we like, divide the topic of instrumental reason into three parts: first, instrumental beliefs – how agents rate the efficacy of means to their ends. Second, instrumental reasoning – the temporally extended process in which agents perform various subsidiary actions based on their ends and their instrumental beliefs. Third, instrumental principles – the various (potentially) normative principles which evaluate the goodness (and possibility) of both instrumental reasoning and instrumental beliefs. Accounting for all three is an essential task for the philosophy of action – a task which this chapter has hardly accomplished. Rather, I've proposed a first step in that overall project, an account of instrumental beliefs that will hopefully yield insights into the nature of instrumental reason generally. But completing that task is a project for another day.

4.0: PRACTICAL KNOWLEDGE

4.1: PRACTICAL KNOWLEDGE IS THE GROUND OF WHAT IT UNDERSTANDS

In the previous two chapters, I argued that a specific kind of representational state – a state which represents itself as the ground of its object – was both necessary and sufficient to account for first, practical categories, including deviance, and second, instrumental reason. Now I will argue that such representations are necessary and sufficient to account for *practical knowledge*: the knowledge that agents typically have of what they are intentionally doing.[125] The argument will come in three parts. First, I will discuss the relation that obtains in typical cases between practical knowledge and action; I will defend the ancient view that practical knowledge is the ground of what it understands.[126] Second, I will argue that the best account of this fact is that intentions are representational states which represent themselves as the grounds of action. When such representations amount to knowledge, then they are in fact the grounds of action. Third, I will apply these ideas to the epistemology of practical knowledge. I will vindicate a thought derived from David Velleman: forming a representation on the expectation that it will be true once formed is an epistemically rational method of representation formation generally; this applies to both beliefs and intentions. Contra Rae Langton and others, this principle does not entail that simple wishful thinking is rational. Thus, a state which represents itself as the ground of its object can amount to genuine knowledge.

I begin with a specific and hopefully uncontroversial claim about practical knowledge: typically, when an agent X is intentionally doing A, X knows "I am doing A." This knowledge has several characteristic features.

---

[125] Use of the phrase "practical knowledge" might suggest adherence to the Anscombean thought that practical knowledge is nonobservational, and perhaps to various of Anscombe's own theses about nonobservational knowledge, e.g. that our knowledge of the position of our limbs is such. But I reject the latter claim, and think the former should not be presumed. At this point, I merely assume that we typically know what we are doing intentionally, and call this practical knowledge.

[126] The phrase is derived from Thomas Aquinas via Elizabeth Anscombe. In fact, Aquinas is in turn adapting an Aristotelian doctrine: that the craftsman's knowledge is the cause of his crafty actions. And Aristotle is of course developing a thought derived from Plato (Kevin Falvey discusses this historical lineage (unpublished §1). In using the phrase, therefore, I don't mean to be endorsing any one of these philosophers' particular developments of it; rather, the phrase is a slogan that can be developed in a variety of ways.

First, it is first-personal: X knows not merely "an A-ing is occurring," but "*I* am doing A." Second, this knowledge can be of some public action, something that someone else might see X doing.[127] Thus, X knows not merely "I am trying to do A," but "I am *doing* A." Third, such knowledge is not restricted to some particular kind of A: tryings, or generic actions, or basic actions, or actions which can only be done intentionally (e.g. promising). Whatever is doable intentionally is a potential object of such knowledge. Lastly, whatever practical knowledge is, it is *knowledge*: justified, true, and non-Gettiered. We might therefore define a new term: "practical representation" – whatever mental state is the analogue of belief; such a state would be like practical knowledge, but unjustified, or untrue, or perhaps Gettiered. I will return to the idea of practical representation (on my view, an intention) in §4.2; for the moment, I stick to the topic of practical knowledge.

Some might balk immediately at this choice of topic. One might say: it is a specifically Anscombean commitment that we know our actions nonobservationally; this is not a neutral starting point. Another might say: practical knowledge is not fundamentally about knowing things in the world, but rather about knowing one's intentions. I respond: both of these claims are compatible with what I have said. I wish to begin by focusing on a specific kind of knowledge under a very generic description: when X is doing A intentionally, typically (but not always) X knows "I am doing A." I do not presuppose that this knowledge is nonobservational, nor that it is the most basic kind of practical knowledge. Instead, I suggest that we can come to decide these other questions by first getting straight on this question: what is the connection between X's knowledge "I am doing A" and the fact that X is doing A?[128]

I defend the ancient view: practical knowledge is the cause of what it understands. Or, since I hold that action can involve relations of non-causal dependence (as discussed in Chapter One §1.2.3): intentional action depends on practical knowledge, or again: practical knowledge is the ground of what it understands. There are a number of opposing views. First, one might hold that practical knowledge and intentional action are identical; there are therefore no relations of dependence between them. Second, one might hold that intentional action grounds practical knowledge. Third, one might hold that they are both grounded in a common cause, e.g. intention. Fourth, one might hold that they ground each other. And fifth, one might hold that there are no dependence relationships

[127] Not all cases of action are like this: consider mental actions, such as performing calculations. I focus on the case of public actions first, but the account I arrive at is meant to apply to actions generally. Thanks to Raja Rosenhagen for emphasizing this point.

[128] In formulating the question in this way I take myself to be following John McDowell (unpublished pp. 9-10).

between them. This fifth option I will ignore, since it simply can't explain why we typically know what we are doing. The fourth option I assume suffers from the defects of the second option, and so I won't give an independent discussion of it. The remaining three possibilities I discuss in turn.

4.1.1: Against identity theories

Perhaps practical knowledge is identical with intentional action. This truth may stand on its own, or it may follow from a more general truth: practical representation, knowledge or not, is identical with intentional action. Against the general thought, I will argue that there can be practical representation without intentional action. Against the specific thought, I will argue that there can be intentional action without practical knowledge.

4.1.1.1: Practical representation is not identical with intentional action. One might first try a simple counterexample: a man intends to bake a loaf of bread. But unbeknownst to him, there is a bomb attached to his stove; as soon as he begins to preheat the oven, both he, his dough, and his house will be reduced to ashes. So while he believes that he is baking a loaf of bread, he is not. But, as Anscombe remarks,

> "A man can *be doing* something which he nevertheless does not *do*, if it is some process or enterprise which it takes time to complete and of which therefore, if it is cut short at any time, we may say that he *was doing* it, but *did not do* it."[129]

So while it is true that the man in question never finishes baking a loaf, he was still certainly in the process of doing so, and so we can say: he was baking a loaf of bread.[130] On the other hand, Anscombe also writes,

> "Sometimes, jokingly, we are pleased to say of a man 'He is doing such-and-such' when he manifestly is not. E.g. 'He is replenishing the water-supply', when this is not happening because, as we can see but he cannot, the water is pouring out of a hole in the pipe on the way to the cistern. And in the same way we make speak of some rather doubtful or remote objective, e.g. 'He is proving Fermat's last theorem'; or again one might say of a madman 'He is leading his victorious armies'."[131]

So Anscombe at least holds that while there are some cases where doing A is compatible with various failures to complete that action, she also holds that there are cases where agents are manifestly not doing what they intend to be doing.

So let us consider a more detailed example. One person says "I am breaking this egg;" I happen to know that what he is holding is not an egg at all, but a game piece from an obscure board game that I happened to have

[129] Anscombe 1957 §23.
[130] A number of authors have followed Anscombe in this thought, including Kevin Falvey (2000 p. 25), Sarah Paul (2009a pp. 16-17), and Michael Thompson (2008 pp. 134-136).
[131] Anscombe 1957 §23.

stored in an egg carton: he is manifestly not breaking "this egg." My opponent might respond: yes, he is breaking an egg, just not yet: in a minute, he'll realize the awful trick I've pulled on him, retrieve a real egg, and break that.

I suggest that this is an "an-this-that" case. Consider: my friend intends to break an egg; it doesn't much matter which, since they're all basically the same. So he picks an egg: and now he intends not just to break an egg, but to break this egg. So we might say: he intends to break this egg in order to break an egg. Unfortunately, he's mistaken: this is not an egg, and he can't break it. So he picks another egg (maybe checking a little more closely this time), and now we can say: he is breaking that egg in order to break an egg. And, moments later, he does just that.

On my opponent's reading, there is one continuous action here: egg-breaking. And this action persists through the various obstacles I have set for it. I suggest that we can be more precise. There are three intentions (break an egg, break this egg, break that egg) and two actions (break an egg, break that egg). Throughout the process, we properly affirm of my friend: he is breaking an egg. But it would be a mistake to infer from this that we must also affirm "He is breaking this egg" when in fact "this" is not even an egg. The impression that there is but a single action here is due to the fact that this is an an-this-that case: an indeterminate end subserved by two successive determinations of it.

In this case, then, my friend has a number of practical representations: that he is breaking an egg, that he is breaking this egg, that he is breaking that egg - and probably many more, insofar as doing each of the above involves yet further actions. While my friend's representation "I am breaking an egg" is accurate throughout the process, this does not entail that his further representation "I am breaking this egg" is also accurate. So this case involves a practical representation without a corresponding intentional action: practical representation and intentional action are therefore not identical.

4.1.1.2: Practical knowledge is not identical with intentional action. Consider an inversion of the above case: LB has been told, repeatedly, by reliable informants, that all the flour in the house has been replaced with a mixture of sawdust and powdered aluminum (for color). And indeed on opening the bag, she does catch a whiff of cut wood. Nevertheless, LB's overwhelming desire for cake leads to a bout of wishful thinking: she holds fast to the thought that what is in the bag of flour is in fact flour. As it happens, she is correct, but she is totally unwarranted in so thinking. And so she goes on baking a cake. In this case, I suggest that LB's practical representation "I am mixing flour into the batter" is totally unwarranted: she should not think as she does. As I said: knowledge, whatever it is, is both factive and *justified*. So LB cannot have practical knowledge of this action. All the same, she really is

intentionally mixing flour into the batter – her understanding of her situation is in fact entirely correct; this is hardly a case of deviant causation. So we have a case of intentional action without practical knowledge: they are therefore not identical.

4.1.2: Against the perceptual model

As noted earlier, one answer to the question "How are practical knowledge and intentional action" related is: the action grounds the knowledge. That is, we have some kind of perceptual awareness of what we are doing. But it's not clear how perception could ever deliver the right kind of knowledge. As I noted at the beginning, practical knowledge is first-personal: X knows "*I* am doing A," not merely "an A-ing is happening" or "someone is doing A." The question therefore is: how can perception deliver this kind of knowledge? There are two answers, neither satisfactory: either perception itself can deliver the relevant kind of first-personal knowledge, or practical knowledge is inferred from perception and a first-personal premise.

I'll first consider the idea that perception can directly deliver first-person practical knowledge. A first pass: perception delivers first-personal knowledge by being perception of oneself. But this faces well-known difficulties: if I see myself in a mirror, and do not realize I am looking in a mirror, the mere fact that I am making judgments about myself does not show that I am making first-person judgments. So one might suggest that it is perception along the right pathway - I'll grant for the sake of argument that proprioception involves no possibility of error through misidentification. Even given this, mere unaided proprioception cannot deliver practical knowledge: I might know by proprioception that my arm is moving, even when this is because someone else is moving it, or it is moving due to a seizure, or electrodes implanted in my arm actuate my muscles in exactly the way I might do intentionally. In any of these cases, nobody would claim "I am moving my arm": "My arm is moving" is correct. Mere perception of the movements of my body, even perception which (I'll assume) is guaranteed to be immune to error through misidentification, is not enough to ground practical knowledge.

Second, perhaps we infer from perception and another premise "I am doing A." Here is a simple model of that inference:

> X knows "X is moving her arm"
> X knows "I am X"
> On that basis, X knows "I am moving my arm"

Perception delivers the first premise. The obvious question is: where does the knowledge "I am X" come from? How is it that I know that this body is mine? We can perhaps answer this question by looking at what sort of evidence would convince me that I needed to update my understanding of which body is mine: if, suddenly, when I tried to drink the glass of water on the table, it was your hand that moved for it, while mine remained entirely still. If this change were systematic and enduring - if my intentions actuated your muscles - that would be evidence that the movements of your body were my actions, while the movements of mine were not. But of course this story involved some *prior ability to tell which actions were mine*. Put simply, the perceptual model has it backwards: we do not know that our actions are our own because they issue from our bodies; rather, we know our bodies are our own because they are the seat of our actions. And this is only possible if we have some practical knowledge of our actions prior to perceiving them.

So of course it is often true that we have perceptual knowledge of our actions; I do not dispute this. But such knowledge cannot be the basis of our first-person knowledge of what we are doing intentionally.

4.1.3: Against the inferential model

As noted earlier, the inferential model explains how we can have practical knowledge of our intentional actions by positing a common cause of both: our intentions. Thus, we arrive at practical knowledge as follows:

> X knows "I intend to do A now"
> X knows "If I intend to do A now, I am doing A"
> On that basis, X knows "I am doing A"[132]

As with the perceptual model, the essential question is: where does the "I" come from? That is, we must now ask: how do agents know what they *intend?* And here my original question recurs: what kind of dependence relations obtain between intention and knowledge of intention? There are four possibilities. First, there are no grounding relations, because intention and representation of intention are identical. Second, intentions ground knowledge of intention. Third, knowledge of intention grounds intention. Fourth, there is a common ground of both.[133]

What I wish to establish is that intention is identical with representation of intention. In particular, intentions represent themselves as the ground of action; I discuss how this content is simultaneously a representation of intention and of action §4.2.2. I arrive at this conclusion via elimination. I will first argue against the idea that

[132] One can also reframe the inference to concern "I will do A"; not much rides on this.
[133] I omit two views: a) that they are distinct but there are no grounding relations and b) that they ground each other, for the same reasons as discussed in §4.1.

intentions ground knowledge of intention. This "inner sense" view cannot explain why we have a sense of control over our actions, but not a sense of control over other effects of mental states known via inner sense. I will then consider the view, developed by Sarah Paul, that intention and knowledge of intention have a common ground. This view, I suggest, collapses into the view that knowledge of intention grounds intention, which in turn collapses into the view that intentions representation of intention is identical with intention, i.e. the view that intentions represent themselves as grounds of action. And this vindicates my initial claim: practical knowledge is the cause of what it understands.

4.1.3.1: Against inner sense. To put the inner sense view metaphorically: we look into the theater of our minds, see our intentions at work, and thus know what we intend. Unfortunately, it is unclear how this view could deliver *essentially* first-personal knowledge. Essentially first-personal knowledge is knowledge immune to "error through misidentification": there's no possibility of my self-knowledge being attuned to somebody else's self - of mistaking your intentions for my own. Here is an example of such an error in a perceptual case: "That is a beautiful animal," I think, looking at John's cow. "And it is Jane's, too," I foolishly believe, "so Jane has a beautiful animal." Here, my conclusion has been arrived at by misidentifying the object of my initial demonstrative belief.

The trouble with inner sense is that it allows for exactly the same mistake when it comes to intentions. I look inside my mind, and see an intention to do A; I believe "That's an intention to do A – and look, it's inside my mind! So I intend to do A." What is odd here is that it seems, as far as inner sense goes, that I might misidentify the intention – I might think that I've seen one of Miriam's intentions, falsely believing that I've recently developed the ability to direct my inner sense at (so to speak) someone else's inner; I might actually see one of Miriam's, and think it mine. Now, of course these narratives are nonsense; that much is clear. The problem is that the inner sense view cannot account for this fact.

Of course, there are responses. Sydney Shoemaker,[134] in response to a similar argument from Richard Moran,[135] points out that there may simply be brute metaphysical necessities, and that one such necessity might be: inner sense only ever points at one's own mind. Now, let us accept that there are brute necessities, at least for the sake of argument. The trouble is that the posited necessity is not metaphysical, but causal. Perception is a capacity for being *affected* by some further thing; my perceptions (or perhaps: my perceptual judgments) are not identical

[134] Sydney Shoemaker seems to suggest this reply to Moran (2003 p. 392).
[135] Richard Moran characterizes this view as "a picture of self-knowledge as a kind of mind-reading as applied to oneself, a faculty that happens to be aimed in one direction rather than another." (2001 p. 91).

with their objects, but caused by them. And causal claims are by their nature contingent. If my beliefs about my intentions are *caused* by those intentions, then it must be possible for such beliefs to be caused by others' intentions. And so we are left with the possibility of errors through misidentification.

And nothing is gained by adding that our own intentions have some special perceptual quality that invariably mark them out as our own – they are not, so to speak, colored first-person-indexical-wise. But assume that they were. Again, the trouble is that perception is an unreliable faculty; occasionally I look at a green object and think that it's blue. So even allowing such a fanciful property, I might nevertheless be subject to errors of misidentification when Miriam's intentions take on that special *I*-glow. And again, the inner sense view cannot explain the fact that this is impossible.

The basic argument here parallels that against the perceptual view of knowledge of action: in neither case do we arrive at a satisfactory explanation of the essentially first-personal character of the knowledge in question. And as before, perhaps there is such a thing as inner sense – a sort of "mind reading as applied to oneself."[136] Nothing I've said shows that such a perceptual faculty is impossible. But it cannot be the source of our self-knowledge of intentions.

4.1.3.2: Against a common cause. Paul, following Bratman, holds that intentions are "distinct practical attitudes," constituted by their world-to-mind direction of fit and functional/dispositional role in practical reasoning.[137] But she recognizes that it is a bit of a puzzle why we should have self-knowledge of something like that:

> "The challenge of investigating self-knowledge is to explain how a subject can come to know in a specially first-personal authoritative way of a given mental state that he is in that state, where this is at least a matter of making a true self-ascription of that mental state. But given the foregoing understanding of what mental states are—diachronic functional states with unique dispositional roles and vague, controversial boundary conditions—this feat begins to look quite extraordinary, and the blanket approach to self-knowledge appears inadequate."[138]

Paul's solution to the puzzle is to introduce the idea of "decision," in the following sense:

> "A decision is paradigmatically a conscious mental act. This is as distinguished from intention, which is a diachronic mental state and plausibly one that does not essentially require any conscious activity on the part of the agent. A decision is a discrete event rather than a state, and one with respect to which we are normally active. It is the act of determining one's will, of settling the question for oneself about what to do by arriving at an answer."[139]

Moreover, a decision to Φ normally causes an intention to Φ:

---

[136] Moran 2001 p. 91
[137] Paul 2012 pp. 328-329.
[138] Paul 2012 p. 329
[139] Paul 2012 p. 336

> "I take it that if deciding to Φ is ordinarily sufficient for coming to be in the state of intending to Φ, the self-conscious making of that decision justifies the self-ascription of the relevant intention."[140]

and

> "...part of what defines the state of intention is the fact that intentions are paradigmatically formed by making a decision about what to do."[141]

Thus decisions are simultaneously the cause of intentions and knowledge of intentions. So far, so good. What I want to focus on is this notion of decision: what exactly does Paul have in mind? She writes:

> "The thought by which one undertakes to direct one's agency toward Φ-ing is a performative type of thought, self-conscious and self-constituting: to take a conscious thought of the form 'It is settled, I shall Φ' to be a decision just is to make a decision," pp. 338-339

So consider Paul's specification of the content of a decision: "It is settled, I shall Φ." By "settled" Paul does not mean normatively settled by Reason - a decision is not a recognition that rationality compels one to Φ. As Paul herself notes, we can decide to do things we recognize are wrong, or we can decide to do one of many equally good alternatives.[142] So Paul must mean *causally* settled.[143] Causally settled by what? By the very decision in question. So we might instead specify the content of a decision as follows: "This very decision will cause me to intend to Φ."

We might model this causal system as:

[140] Paul 2012 p. 339
[141] Paul 2012 p. 339
[142] Paul 2012 p. 334.
[143] Paul suggests that she means that such decisions settle matters both causally and normatively, this latter in the sense that these decisions represent themselves as being "anchors" in further reasoning about what to do. When I decide to do A, I need to start reasoning from that decision, or I am rationally criticizable. (Paul, personal communication.) I suggest that, first, it is possible to explain this normative thesis using the fact that decisions represent themselves as causally settling matters: if the agent doesn't reason from them in the appropriate way, then the decision is false – it simply hasn't settled things – and the agent is incoherent. I suggest how such an argument might run in Chapter Three (§3.5.4). Second, for my purposes here, all that matters is that the content of decision includes that they causally settle things; whether they also represent themselves as normatively settling matters is not, at the moment, essential. Thanks to Sarah Paul for helpful discussion of these issues.

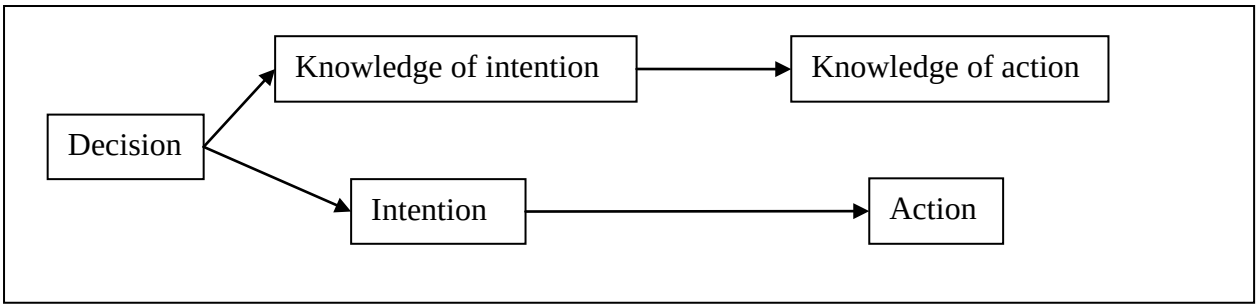

Figure 3

What I will argue is that once we've gone this far, we might as well just say: intentions are representations of the form "This very representation will ground A."

The first question is: why does Paul insist that decisions are a distinct mental act, rather than a state? The point, I take it, of making decisions "paradigmatically conscious acts" is to ensure that they're epistemically accessible, i.e. that the following reasoning makes sense:[144]

> X decides "I will do A," i.e. "This very decision will cause me to intend to do A"
> On that basis, X believes "I intend to do A"

But why not just skip that somewhat repetitive reasoning, and simply say that "decisions" are not conscious mental events, but are already representational states?[145] If we begin with a representational state of the form "This very state will cause me to intend to do A," then there's no need to posit some further step from decision to representation of intention: decisions already represent intention.[146] We don't need a detour through acts which exist solely to rationalize states.[147] We can therefore revise the model as:

[144] Paul suggests something along these lines (2009a p. 13).
[145] A point of clarification: in my view, decisions are in fact events, namely, the formation of an intention (or perhaps some subset of those, perhaps the ones involving some degree of conscious deliberation). What I mean to suggest in the text is not an analysis of the concept of decision; rather, I wish to argue that Paul's technical notion of "decision" is better replaced by a likewise technical notion of intention, a representational state rather than an event.
[146] I defend this move in more detail later in this chapter (§4.2.2).
[147] Paul suggests that the distinction between decisions and intentions may be necessary to allow that some decisions fail to lead to intentions. But she also recognizes that even without that distinction one may account for the phenomena by noting that some intentions may fail to lead to further instrumental reasoning and action. Paul, personal communication. I discuss these points in connection with the topic of unconscious intentions in Chapter Five (§5.4).

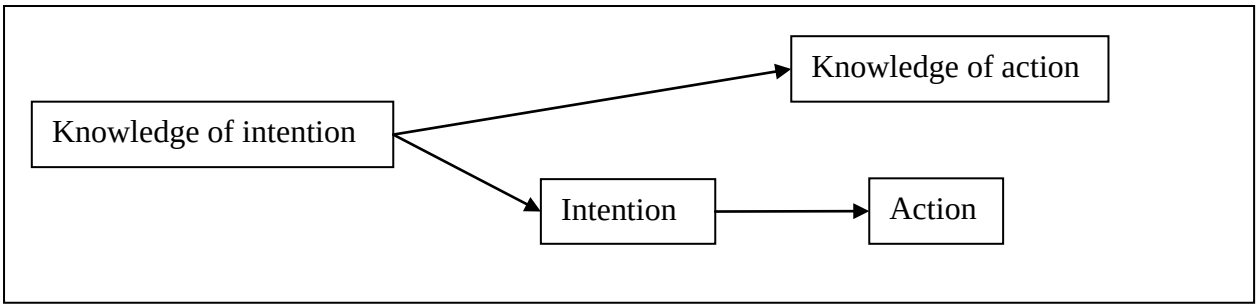

Figure 4

So we have mental states which represent themselves as causes of intention. The next question is: why not say that there are mental states which represent themselves as causes of action? Once we've let this kind of knowledge of intention in the door, why not extend that account to knowledge of action? Paul might answer: we need the detour through intention to account for some other aspect of intentional action, such as the practical categories, or instrumental reason. But, as I argued in Chapters Two and Three, this isn't necessary: states which represent themselves as the ground of their objects can do all the necessary work. Paul has also suggested that we need a separate notion of intention to account for the occasional unconscious intention.[148] Following Matt Boyle, I suggest that both Paul and I can avail ourselves of the same account of unconscious intentions.[149] On both theories, the explanation of unconscious intentions is the same: a failure of normal inferential processes which normally connect intentions with the rest of the agent's mental states. Whether or not intention is a distinct practical attitude does not affect our ability to give this kind of explanation. I discuss this issue in more detail in Chapter Five (§5.4). We can thus revise the model again:

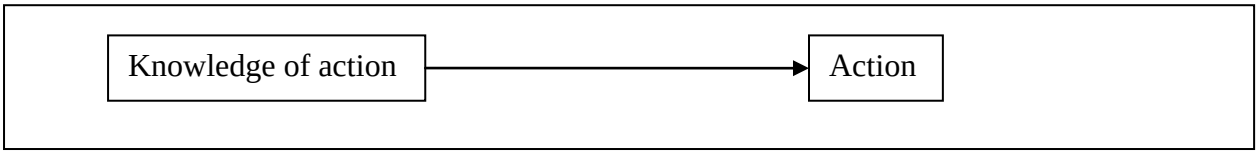

Figure 5

[148] Paul 2009a p. 5.
[149] Boyle makes this point in his 2011 (p. 229).

In brief: once we introduce the idea of decisions, then we might as well just say: intentions are representations of the form "This very representation will ground A." And, since this state represents itself as existing, intentions are identical with representations of intention. Or, equivalently, practical knowledge is the ground of what it understands.

4.1.4: Summary of the argument so far

I began with this question: what is the relation between action and knowledge of action? There are five possibilities: representation of action is identical with action, knowledge grounds action, action grounds knowledge, knowledge and action have a common ground, intention, or neither depends on the other. The last can be ruled out immediately: it leaves it completely inexplicable why we typically know what we are doing. I then argued that representation of action is not identical with action, since there can be practical representation without intentional action, and intentional action without practical knowledge. I then argued that the view that action grounds knowledge faces a serious problem: it cannot account for the essentially first-personal nature of practical knowledge. I then argued against the view that action and knowledge have a common cause in intention. On this last view, knowledge of action is inferential: we first know our intentions, and infer our actions on that basis. This raises the question: how do we know our intentions? As before, there are several possibilities: representation of intention is identical with intention, knowledge causes intention, intention causes knowledge, intention and knowledge have a common cause, or they have no causal connection. The last view can be ruled out for the same reason as before: it can't explain why we typically know what we intend. I then argued that the view that intentions cause knowledge of intention, like the analogous view about action, cannot account for the *essentially* first-personal character of such knowledge. I then argue that the common cause view, as developed by Sarah Paul, collapses into the view that knowledge of intention causes intention, which in turn collapses into the view that intention is identical with knowledge of intention – that is, my own view, that intentions represent themselves as the ground of action. So I conclude: practical knowledge is the ground of what it understands.

4.2: PRACTICAL REPRESENTATION REPRESENTS ITSELF AS THE GROUND OF ITS OBJECT

In the previous section, I focused on "the good cases," cases where X is doing A, and *knows* "I am doing A." And I argued that in such cases, X's knowledge is the ground of X's action. And the simplest explanation of this fact is that practical representation represents itself as the ground of action. In the good cases, the agent's practical representation is knowledge, and therefore true and justified: and thus the agent's practical representation is the ground (because it is true) of what it understands (because it is justified). In bad cases, we have a representation with the same content, but it in one way or another does not amount to knowledge: perhaps it is false but justified, and therefore an intelligible failure, or true but unjustified, and therefore a lucky success, or both, and a straightforward failure. In any case, the content of practical representation is: "This very representation will ground A." There are a number of questions one might have about this formulation. First, how does it apply to cases where success is unknown or unlikely? Second, how does this new formulation correspond to my prior focus, "I *am doing* A"? Third, how does this formulation connect representation of action and representation of intention?[150]

4.2.1: Cases where success is unlikely

What of Davidson's notorious carbon copier, who intends to make ten carbon copies, but is not sure that he's actually succeeding? Here, I suggest simply relying on an analogous problem in mere causation: *probabilistic causation.* And here, as Woodward (following Cartwright) suggests, we can give a precise definition of the extension of a "full causation" framework to a "partial causation" framework.[151] Considering two binary variables, {A,~A} and {B,~B}, A is a probabilistic cause of B just in case an intervention from ~A to A increases the likelihood of B. *This does not mean that B is particularly likely to happen*. For instance, the intervention might raise the likelihood of B from .01 to .02. Here, even given A, the reasonable expectation is ~B. Similarly with practical representation: in cases where success at A is unlikely, the agent may yet think "This very representation will ground A," but assign it a relatively low credence.[152] To return to the carbon copier: he presumably takes his intention to make ten carbon copies to raise the chances of his actually making ten carbon copies, even if he knows it does not guarantee it.

---

[150] The following ideas, especially those in §§4.2.2-3, are heavily indebted to Richard Moran's (2001) as well as Matt Boyle's adaptation of those ideas in his (2011).
[151] cf. Woodward 2003 pp. 61ff.
[152] This does not entail that one must assign a high credence to "This very representation will cause ~A;" this is an important feature of the logic of representation-centered representations as opposed to agent-centered representations.

Kieran Setiya advances a similar proposal in his "Practical Knowledge."[153] Sarah Paul, in her "Intention, Belief, and Wishful Thinking" objects: it might be that some autonomic process is a more reliable cause of the intended outcome than one's intention. So intending the outcome actually lowers its likelihood.[154] Setiya responds that this depends on what the relevant counterfactual situation is; to speak causally, if we fix the autonomic process to "off," then the intention does raise the probability of the outcome.[155] While Setiya is correct, it seems preferable to avoid any reliance on intuitions about which counterfactual situation is "salient" or "relevant" here. Instead, I suggest that a precise Woodward-style definition of ground, plus a precise specification of the content of the agent's intention, can solve this problem. The precise definition is this: the agent believes that an intervention on their intention from "off" to "on" will, holding fixed at their actual values the values of variables not on a direct path from the intention to its object, increase the probability of the outcome.

So what is the precise content of the agent's intention? If the agent's intention is to breathe regularly, via whatever means, then conscious control over their breathing seems like a terrible plan.[156] But this intention would entail *not* exercising conscious control, i.e. would be a *cause* of not exercising conscious control, and would therefore increase the probability of actually breathing regularly.[157] That is, given an intention to breathe regularly, and the knowledge that conscious control over breathing actually decreases regularity, I might decide to distract myself with a good book *in order to* breathe regularly. In this way, my intention to breathe regularly is in fact a cause of me not exercising conscious control over my breathing, and thus a cause of regular breathing.

On the other hand, if the agent's intention is to breathe regularly via conscious control, then it makes perfect sense to exercise conscious control, and again, the intention increases the probability of the outcome (or at least, the agent takes it to). Consider an analogy to with a case of instrumental reasoning: let's say that we are watching X play a game of ultimate frisbee. And X is at all times holding their hands next to their shoulders, in the fashion of a

---

[153] Setiya 2008 p. 391.
[154] Paul 2009b p. 553.
[155] Setiya 2009 p. 130.
[156] I here assume that an intention to breathe regularly is not ipso facto exercising conscious control over one's breathing. As I argued in Chapter One (§1.3.1), we should distinguish intentions as representational states from anything consciously occurrent, involving attention, etc. However, if this is denied, then in the situation described, I would simply say: one can't (rationally, consistently) intend to breathe regularly, since any such intention will decrease one's chances of success.
[157] This is perhaps a bit tricky, since it involves causation via overdetermination of an omission. That is, the intention causes something to not happen (conscious control) which would have not happened even without the intention (overdetermination). I suggest however that the agent can still rationally take their intention to increase the likelihood of breathing regularly, insofar as it makes their not exercising conscious control more secure against possible disruptions (to put it loosely). For a more detailed discussion of causation via overdetermination and omission, see Woodward 2003 Chapter Two §§7-8.

Tyrannosaurus rex. We ask why; X responds "In order to score a point." And if we were to point out that holding one's hands like that actually makes it very difficult to score a point, and X were to respond, "Yes, that's right, but still, I am holding my hands like this in order to score a point," then we would start thinking that X was putting us on. On the other hand, if X initially responded "In order to score a dinosaur point," (which just means: to score a point while holding one's hands in the fashion of some dinosaur), then X's action would make perfect sense: even though it decreases the chances of scoring a point, it increases the chances of scoring a dinosaur point. These facts about instrumental reason would be inexplicable if it were perfectly fine to have an intention one takes to lower the chances of success: the character of instrumental reasoning is derived from the character of the ends it serves, and so if instrumental reasoning must increase one's chances of success, this is simply because the goal of such reasoning, the intention it serves, is about increasing one's chances of success. Thus, if X appears to take their intention to do A to lower the chances of A, then we have misspecified the content of X's intention.

4.2.2: Representation of action and representation of intention

Earlier (§4.1.3), I suggested that intentions are simultaneously representations of intention and representations of action. This seems a bit mysterious, since in many cases these are just distinct representations. Consider the following passage from Uriah Kriegel, in which he expounds a similar view concerning beliefs:

> "...when I self-consciously think that the almond trees are blooming again, I have a thought M_1 with two contents: the primary content is the proposition <The almond trees are blooming again>, whereas the secondary content is something like the proposition <I am herewith thinking that the almond trees are blooming again>."[158]

The basic problem with Kriegel's proposal is that we typically individuate thoughts by their content: if we have two different contents, we have two different thoughts; if we have the same content, we have the same thought. Consider the strangeness of the question "How many beliefs that p does X have at t?" So Kriegel's proposal won't do, insofar as it attributes two quite distinct contents to the same thought.

Instead of having a thought with two distinct contents, I suggest that a single content with several different implications does the trick. As Michael Thompson notes, the single thought "Cato killed Cato" implies many otherwise independent thoughts, e.g. "Cato killed someone" and "Someone killed Cato."[159] So consider the belief "q because of p." Such a belief implies p, q, and that there is a grounding relation between them. Thus, a

[158] Kriegel 2003 p. 126.
[159] Thompson 2008 p. 95

representation of the form "This very state will ground A" implies A, that it itself exists, and that a grounding relation obtains between itself and A.[160] Thus, on my view, the representational content of intention simultaneously includes representation of intention and representation of action. Of course, intention and action are different things: it might be true that X intends to do A, but X will not do A. In such a case, we might say that X knew "I intend to do A," but did not know "I will do A." But these different facts were, in X, represented by one and the same thought.

4.2.3: The present progressive and the simple future

I say that the content of practical cognition is: "This very representation will ground A." But for most of this chapter, I've focused on present-progressive thoughts, i.e. "I am doing A." What I suggest here is that the agent's practical representation includes both of these thoughts. This can seem a bit mysterious, since the distinction between the present progressive ("I am doing A") and the future perfect ("I will do A") is often taken to be quite basic. But as with intention and action, I suggest that intentions simultaneously represents these different contents.

In order to show this, however, I need to take a slight detour, and discuss the details of the connection between the present progressive and the future perfect. Sebastian Rödl remarks:

> "A movement [i.e. the present progressive] falls under a form, and this form sets the measure by which it is to be decided whether the movement has reached completion [i.e. the simple future]."[161]

and also

> "...a movement's form, and thus its end, which the form designates, is present in the movement as it is progressing."[162]

So consider the thought "This very representation will ground A." That is, this thought represents itself as the cause of the completion, the perfection, of A-ing. So it implies the thought "I will do A," simple future. But it also represents itself as the actual present ground of that "will do A," and thus also implies "I am doing A." Indeed, this thought provides the rule that guides the agent's instrumental reasoning: in this way, this thought is both the measure and the ground of the movement.[163] Of course, these different components of the content can be falsified separately: it might be true that I am doing A, even when I will not do A. But practical cognition encompasses both thoughts simultaneously.

---

[160] I discuss the implications of this view for the possibility of unconscious intentions in Chapter Five (§5.4).
[161] Rödl 2005 p. 172
[162] Rödl 2005 p. 171

4.3: THE EPISTEMOLOGY OF PRACTICAL REPRESENTATION

In the first part of this chapter, I argued that practical *knowledge* is the cause of what it understands. In the second part, I drew a conclusion about the *content* of practical representation generally: it *represents itself* as the cause of its object. In this part, I will explain what this account of the content of practical representation means for its *epistemology*. That is, I will explain *how* representations with such a content could ever amount to knowledge.

To begin, one might be skeptical about the idea of practical knowledge for very general reasons. I say it is knowledge not had by observation of the action in question, even though the action is the sort of thing which can be observed by the agent or anyone at all.[164] But the basis of practical knowledge is therefore compatible with the action not occurring, if the agent fails to pull it off. So one might argue: the agent doesn't *really know* what they are doing until they see it happening; before then, they are just making an educated guess. I respond: this argument is in fact quite general. Consider the following dialogue:

> **GA**: The sun will set tonight.
> **The Philosopher**: You might think that at first - but all your evidence hardly *guarantees* that the sun will set tonight. So you won't really know that the sun will set until you see it happen.
> **GA**: But by that logic, I also don't know about the past. For all my *current* evidence is logically consistent with the sun not having set last night either.
> **The Philosopher**: That's right! In fact, even if it appears that the sun is setting, you don't *know* that the sun is setting, since appearances can also mislead.
> **GA**: So I should properly say that I can only know the contents of my own mind?
> **The Philosopher**: Yes, that's exactly right!
> **GA**: But this is a mad account. Instead: we can know about the world via perception, the past via memory, the future via prediction. Evidence need not *logically guarantee* the truth of the conclusion.

So set this very general worry aside: general skepticism about knowledge of the external world is no objection to practical knowledge in particular.

---

[163] It is possible that conditional intentions ("I will do A, if p") do not fit this model, since they do not yet imply either "I will do A" or "I am doing A." But this does not undermine the general picture of intentions I have offered here, which is quite compatible with it. That is, such intentions still represent themselves as the ground of their object, albeit conditionally. And this explains why conditional intentions do not underwrite progressive action attributions. (It is possible that some intentions for the future are best understood as conditional intentions (i.e. with a temporal condition); this would explain why some intentions for the future do not underwrite present-tense progressive action attributions.)

[164] Cf. Anscombe 1957 pp. 51-52. Moran calls this "the most basic difficulty" with practical knowledge (2004 p. 48).

What I take to be more serious is a longstanding challenge to cognitivist theories of intention developed by H.P. Grice, Rae Langton, and Sarah Paul.[165] Grice writes:

> "If we ask why (assuming the theory to be correct) I hold a belief that (say) I shall go abroad, the only possible answer is that I believe this because, for example, it is something that I should particularly like to be the case or something that I think ought to be the case. This is not only common form, it is regarded by us as entirely reasonable. So, to put it crudely, the theory represents having an intention as being a case of licensed wishful thinking."[166]

Grice's argument is brief, but straightforward. On my view, intentions are representational states, correct only if they accurately represent the world. But they are formed not on the basis of an expectation that the world is a certain way, but a desire that it should be that way. So they cannot be epistemically justified, even if perhaps they are practically justified.

The basic challenge is simple: I must show that forming intentions can accord with some epistemic principle which does not in turn license obviously irrational cases of wishful thinking. In truth, I think the challenge has already been met by Velleman; his solution just needs a spot of paint here and there. I will rely on a few basic ideas. First, there is a connection between X's expectations about whether a possible representation would be true or false, and whether X is permitted to form that representation. Here's a rough formulation:

> *Expectation of Truth*: If X believes[167] "If I represent p, that representation will be accurate" then prima facie,[168] X is permitted[169] to form the representation that p.[170]

This suggests the following analogous principle:

> *Expectation of Falsity*: If X believes "If I represent p, that representation will be inaccurate" then prima facie, X is not permitted to form the representation that p.[171]

---

[165] Grice 1971 p. 8; Langton 2004; Paul 2009b.
[166] Grice 1971 p. 8.
[167] One might prefer to formulate this principle as "If X *knows*..."; one is welcome to do so, since it should not materially affect the discussion to do follow.
[168] The "prima facie" is necessary to deal with various conditions which would undermine X's entitlement. Indeed, it is my view that such conditions hold in certain wishful-thinking cases, as I will argue in a moment.
[169] Why "permitted" and not "required"? Because, as Velleman shows, there may be cases where one is permitted to form one of many incompatible representations, since any one of them, if formed, would be true (1989 pp. 37-40).
[170] I should note at the outset that there is a close neighbor of this principle which would not license forming intentions, as I've described them; call it *Truth First*. Rather than concerning whether a representation would be accurate if formed, it concerns the propositions referred to by potential representations; if X believes that that proposition is true *already*, then X is permitted (or perhaps obliged) to form the representation. The slightly circuitous nature of my formulation of *Truth First* is needed to handle cases of indexical thoughts; if I just stated the principle as "If X believes p is true, then X is obligated to form the belief that p," then if we substitute an indexical content for "p," then that thought winds up indexed to different situations (roughly, to the situation prior to forming the belief, and to the situation after doing so), and the principle is invalid. That said: as far as I can tell, the *only* disagreement between *Truth First* and *Expectation of Accuracy* concerns cases of forming intentions, as I've described them; there seems to be no common ground by which one could adjudicate these two principles. Thus, the most I can show here is that a plausible epistemic principle (*Expectation of Accuracy*) licenses forming intentions, not that *every* plausible epistemic principle does so.

I'm not, however, willing to endorse the analogous biconditional. There might be cases where X is permitted to form a representation despite having no expectation one way or the other. But settling this question is, I hope, irrelevant to the topic of wishful thinking. Second: I've said that intentions, like beliefs, are a kind of representational state. Both are correct only if accurate; they are distinguished by a formal feature of their content: intentions represent themselves as the cause of their objects. This suggests an analogous claim about beliefs:

Nature of Belief: Beliefs represent their objects as not causally dependent on themselves.[172]

This principle, while contentious, is at least not ad hoc: it is the natural counterpart to the theory of intentions I've been developing. However, even if one is inclined to deny *Nature of Belief*, the following vaguer principle might be more acceptable:

Function of Belief: beliefs function to represent independently existing reality.

In some sense, *Nature of Belief* is just one way of spelling out *Function of Belief* in precise terms; but there are other roads to take here. However, these other roads are hopefully not relevant here: while I'll rely on *Nature of Belief* in the pages to follow, largely the same points can be made using *Function of Belief* instead; I leave it to the reader to substitute *Function* for *Nature* where they deem appropriate. These claims - *Expectation of Truth*, *Expectation of Falsity*, and *Nature of Belief* - are, I suggest, sufficient to show that a) intentions, as I've defined them, can amount to knowledge, and b) other kinds of wishful thinking are irrational. I'll demonstrate this by consideration of the various cases that lie along the spectrum between simple wishful thinking and instrumental reasoning.

---

[171] This principle requires some complications in "anti-expertise" cases (cf. Egan and Elga 2005), i.e. cases where X knows "If I believe p, that belief will be false" *and* "If I believe ~p, that belief will be false." In such cases, the principle permits neither the belief that p nor the belief that ~p; presumably therefore it commands that X suspend judgment. However, one might complicate the case further, and add that X also knows "If I suspend judgment on p, then p." If we interpret suspending judgment on p as a representation which is inaccurate if p, then the principle will also forbid suspending judgment. If we follow that interpretation, then the principle will, prima facie, forbid every option. But this is not really a problem, I think, since in such a case X is rather caught in an epistemic bind: there *is* no good option. On the other hand, if we don't follow that interpretation (and I think we should not), then X will be forced to suspend judgment, despite having on hand the premises in a deductive inference that p; I would therefore be forced to say that one should not always make deductive inferences from known propositions. But again, given the difficulty of the case, this is perhaps not the worst result. So, in summary, while the principle involves some odd conclusions in anti-expertise cases, these conclusions arise primarily from the difficulty of the case itself, not from any failings of the principle. Thanks to Robert Steel for insightful discussion of these issues.

[172] One might be inclined to make further distinctions: perhaps perceptual beliefs represent themselves as caused by their objects. This shows why I do not employ the simpler claim that beliefs represent themselves as *independent* of their object: that implies that there is no causal connection in either direction. But the example of perceptual beliefs shows why this cannot be true in general.

Consider first simple wishful thinking: X prefers p, and on that basis, believes p. The obvious problem is that X has no reason to expect that her belief will be true; indeed, in typical cases, X has good reason to think that her belief is false. *Expectation of Falsity* suffices to show why X is irrational.

Consider next "post hoc justification." Setiya and Paul read Velleman as endorsing this sort of principle: if X believes "If I form the belief that p, that belief will be justified," then X is permitted to form the belief that p.[173] This principle, however, is subject to some obvious problems. If I expect that I am about to receive some misleading evidence that p, then it is obviously irrational to a) form the belief that p and b) drop the belief that the evidence is misleading. Again, *Expectation of Truth*, being cast in a more objective mode, is not subject to these problems.

So let's consider a more difficult case: the beneficent demon. Langton, adapting an example from Lloyd Humberstone, considers a beneficent demon who ensures, as best it can, that X's beliefs are true.[174] And indeed, let's consider that X knows about this. Still, Langton suggests, it would be irrational for X to form the belief that p on the basis of a preference for p. Here, *Expectation of Truth* is no help: it would, prima facie, permit X to believe that p. However, I suggest that *Nature of Belief* shows that X has incoherent beliefs. On the one hand, X believes that p, and by *Nature of Belief*, this belief represents p as not causally depending on itself. On the other hand, X also believes that there is a causal path that runs from her beliefs via a demon to their objects. So she in some sense has an incoherent set of beliefs: and, I suggest, it is probably (epistemically) irrational to form a belief that you know does not cohere with the rest of your beliefs.

However, Langton's case can be modified. Consider a demon which follows this rule: "if X forms the belief that p, see to it that p, unless that conflicts with a prior application of this rule." In such a case, while there is a causal path from the *formation* of X's belief, once formed, there is no causal connection between X's *belief* and p. So I can't appeal to *Nature of Belief* in just the same way as before: there is no time in which X is in an incoherent *state*. But, to speak somewhat figuratively, if intentions represent themselves as the cause of their objects, then so

[173] E.g. Setiya 2008 p. 398, quoting Velleman 2007 pp. 56-57; Paul 2009b p. 547. Velleman's discussion is unfortunately imperspicuous, but I do not think he actually endorses the principle in question. Rather, the discussion that follows the passage Setiya cites (e.g. Velleman 2007 p. 57 n. 12) makes it clear that he is relying on *Expectation of Truth*, and not the invalid principle that Setiya and Paul take him to be relying on. This is more evident in Velleman's 1989, esp. pp. 38-39.

[174] Langton 2004 p. 257; Humberstone 1992 p. 62.

does the process of practical reasoning; if beliefs represent their objects as not depending on them, then so too does the process of theoretical reasoning. Thus, if X *theoretically reasons* to the belief that p, then that *process* is irrational, since it does not cohere with the rest of X's beliefs.

So now let us consider forming an intention. X believes "If I intend to do A, I'll do A." So *Expectation of Truth* says: X is permitted to intend to do A. And this involves no incoherence: while X does take her intention to be the cause of its object, intentions are unlike beliefs: they represent themselves as the cause of their objects. So forming an intention in this way is epistemically permissible, and involves no incoherent states or processes.[175] Thus, I suggest, such an intention can amount to practical knowledge.[176]

All that said, one might object, as Langton does: but one simply *can't* form representations at will. So this story about intentions can't possibly be correct. She cites Williams' discussion of believing at will:

> "...it is not a contingent fact that I cannot bring it about, just like that, that I believe something...Why is this? One reason is connected with the characteristic of beliefs that they aim at truth. If I could acquire a belief at will, I could acquire it whether it was true or not; moreover, I would know that I could acquire it whether it was true or not. If in full consciousness I could will to acquire a 'belief' irrespective of its truth, it is unclear that before the event I could seriously think of it as a belief, i.e. as something purporting to represent reality."[177]

Now, I am happy to accept that we can rarely believe at will. Yet if "forming a *representation* at will" means "forming a representation on the basis of practical reasons," then it is straightforwardly false. As I noted in Chapter One (§1.3.1), it is a widely-accepted view that intentions are representational states. And they are formed on the basis of practical reasons all the time. This does not mean that I may form intentions by intending to intend, nor does it imply that one can form any arbitrary intention whatsoever. Rather, we characteristically form intentions on the basis of ordinary practical reasoning – there is no mystery here.

Indeed, the account I have developed here offers a nice explanation of when and why we can form representations on the basis of practical reasoning. We human beings have some difficulties engaging in self-consciously irrational reasoning, and as I've just argued, forming a belief on the basis of practical reasons is irrational. Similarly, the epistemic principles I've stated can explain why we cannot just form any arbitrary intention at will. I know that if I were to intend to grow wings and fly *right now*, I would of a certainty fail. So *Expectation*

---

[175] Langton considers what are essentially these points, and concludes "I presume these conditions do not, saliently, improve his beliefs," (2004 p. 258). But she adduces no particular reasons to support her presumption.
[176] Setiya suggests that this is not a sufficient story, since it ignores the role of know-how (2008 pp. 404-406). I think this is a virtue, not a defect, of my account; contra Setiya, I don't think that if X is doing B intentionally, then either X knows how to B, or is doing B by means of some A which X knows how to do. But this is because, as discussed earlier in this chapter, I hold that lottery cases are cases of intentional action.
[177] Williams 1970 p. 148.

of Falsity tells me: don't form that intention. Since forming it would be irrational, it is quite difficult to do so. On the other hand, I could easily form an intention to, say, reach out and move a matchbox – nothing easier.[178]

4.4: SUMMARY AND CONCLUSION

In this chapter, I discussed practical knowledge: the fact that, typically, if X is doing A intentionally, X knows "I am doing A." I began with the question "What dependence relation(s) exist between action and practical knowledge?" Via a process of elimination, I arrived at the conclusion that action depends on practical knowledge. I then argued that the best explanation of this fact is that intentions represent themselves as the ground of their objects. If such representations are formed via sound reasoning, then they amount to knowledge, and thus are the ground of their object. Such reasoning is guided by a valid epistemic principle, roughly "You may form a representation if you expect that it will be accurate." Because of the different character of beliefs and intentions, this principle licenses practical reasoning, but not mere wishful thinking.

[178] cf. Anscombe 1957 p. 52.

5.0: RESPONSE TO OBJECTIONS

5.1: INTRODUCTION

In this chapter, I respond to a number of "big picture" objections to the theory of action I've developed. The objections themselves largely do not depend on the details of my account, but apply to any similar theory. However, in responding to them, I will draw on the details of the particular account I've developed. In general, these objections allege that there is some stretch of intentional action that my theory is unsuited to account for, or, in one case, that it incorrectly classifies some unintentional behavior as intentional.

In §5.2, I address a problem case I'll call "Parfit's Insomniac." The basic objection is that agents can have the representational state I identify with practical thought but not act intentionally. I'll argue that the initial example involves bad reasoning; seen more clearly, my account of intentions can properly distinguish between genuine intentions and mere self-fulfilling prophecies.

I'll then (§5.3) turn to the problem of animals and small children. It is often argued that such creatures lack the relevant conceptual capacities, but act intentionally. I'll argue that the empirical results have been misinterpreted. Animals and children lack the full flowering of the theory of mind, but cognitive explanations of a wide range of behavior require that they possess some partial degree of it.

Next (§5.4) I'll discuss the topic of unconscious intentions, exemplified in a story about Freud's inkwell. The basic objection is that we can act on the basis of intentions of which we are unaware. But on my view, intentions are essentially conscious. Or so one might argue. I'll respond that my theory can avail itself of exactly the same account of unconscious intentions as a distinct-practical-attitude view: they are a breakdown of normal inferential processes.

Lastly (§5.5), I'll turn to the gap between the philosophical theory of action here developed and ordinary practical thought. My own view is that the theory makes explicit what is present but inexplicit in ordinary life. Thus, there is no conflict, metaphysical or otherwise, between my account and ordinary practical thought.

5.2: PARFIT'S INSOMNIAC

> "An insomniac may find that the belief that he is going to stay awake is a self-fulfilling belief, which may lead to his being able to make the self-referential prediction that he will stay awake because of that very prediction..."[179]

According to Gilbert Harman, this example shows that intentions aren't merely self-referential predictions, since the insomniac makes such a prediction but does not intend to stay awake. I respond: the account of intentions I have here developed has the resources to distinguish between the insomniac's thoughts and genuine intentions both in theory and in practice. First, let's break down the chain of reasoning Harman describes:

> The Insomniac believes "I will stay awake tonight"
> The Insomniac believes "That belief is going to cause me to stay awake"
> On that basis, the Insomniac believes "This very belief is going to cause me to stay awake"

As I discussed in Chapter One (§1.3.2), this is bad reasoning.[180] Just because the Insomniac knows he has one mental state with a certain property does not mean that he is required to form a new representation that attributes that property to itself. Indeed, just given those premises, he is not even *permitted* to.[181] The fact that one of my mental states has a certain property is no guarantee that some other different state would also have it.

I take these reflections to show that the account of intention I have here developed has the conceptual resources to mark a distinction between the insomniac's thoughts and genuine intentions. But one might still object that this is mere words; I need to also show how these theoretical differences show up in the Insomniac's actual behavior.[182] Here, I have recourse to the ideas developed in the previous chapters. A thought merely of the form "I will stay awake tonight" cannot ground any of the essential features of intentional action – practical categories, instrumental reasoning, practical knowledge. For instance, one notable feature of the Insomniac is that his belief "I will stay awake" does not in any way commit him to taking further means to ensure that he actually does stay awake; an intention to stay awake, by contrast, would. The Insomniac just lies there, and if he detects the onset of sleep, that's a welcome development, not something he might respond to by brewing some coffee. In other words,

---

[179] Harman 1976 p. 448; the example is due to Parfit. Harman presents a modified version of the example in his 1986 (p. 375); unfortunately, his discussion there is even more confused than the earlier one. I suspect that Rae Langton develops a similar case in her 2004 (p. 13). My response to Parfit, then, also constitutes a response to Langton's objection.

[180] Michael Thompson also suggests that the case may be confused by a failure to distinguish between a representational state and an occurrent thought. In actual cases of insomnia, insomniacs are plagued not by persistent *beliefs* – that would make insomnia a sort of psychotic disorder – but rather by persistent occurrent thoughts, more akin to imaginings than beliefs. I discuss this confusion in more detail in Chapter One (§1.3).

[181] At least, given the epistemology developed in Chapter Four (§4.3).

[182] Thanks to Sarah Paul for pressing me on this point.

the distinction between "I will stay awake" and "This very thought will cause me to stay awake" is not a distinction without a difference; rather, it is a distinction that makes a very important difference to the agent's practical reasoning.

Despite all this one might still press the objection. Let us assume that rather than engaging in the bad reasoning described earlier, the Insomniac simply thinks "This very thought will keep me awake," and it does so, in exactly the way he expects it to. Isn't it intuitive that this is simply not an intention?[183] I suggest that the intuitive force here derives from the thought that the insomniac's insomnia is in some way not up to him. If it is out of his control, then it cannot be an intention of his. Instead of an intention, we have here a compulsion. As Sarah Buss puts it:

> "When an agent forms an intention by deliberating about what to do, her intention is not something that merely happens in and to her. It is something she makes happen, something she plays an active role in producing."[184]

Seen in this light, the essential problem for my theory stemming from Parfit's Insomniac is not a mistaken inference, but a lack of control over his intentions: my theory does not guarantee that agents have such control. The Insomniac is therefore most akin to an addict. But, I suggest, this comparison vindicates my theory rather than undermining it. Addicts may in some sense lack control over their intentions: but, contra Buss, they are intentions all the same. After all, addicted action (that is, action due to addiction) displays the hallmarks of intentional action: it involves a distinction between intended and side effects, it involves instrumental reasoning, and it involves practical knowledge. Now, of course we want to in some way distinguish addicted action from other kinds of intentional action: but this is a distinction within intentional action. So I conclude: the fact that the Insomniac, like the addict, in some unclear sense lacks control over his actions does not show that his action is not intentional.

In summary: the Insomniac does indeed pose a serious challenge for views like mine. It is my hope that the arguments of the preceding chapters show how that challenge can be met.

5.3: ANIMALS AND SMALL CHILDREN

> "The first point is simple: by their very nature, meta-cognitive processes contain an extra layer of representational complexity. A creature that is capable of metarepresenting some of its own cognitive

[183] Thanks to Kieran Setiya for pressing me on this point.
[184] Buss 1999 p. 403. Buss later explicitly denies that addicts act intentionally when they satisfy their addictions (p. 419).

> processes must first, of course, have the wherewithal to undergo the first-order processes in question. Then to this must be added whatever is necessary for the creature to represent, and come to believe, that it is undergoing those events. Put differently, a creature that is capable of thinking about its own thought that P must be capable of representing thoughts, in addition to representing whatever is represented by P.
>
> "The second point is that in the decades that have elapsed since Premack and Woodruff (1978) first raised the question whether chimpanzees have a ‘ theory of mind ’, a general (but admittedly not universal) consensus has emerged that metacognitive processes concerning the thoughts, goals, and likely behavior of others is cognitively extremely demanding (Wellman, 1990; Baron-Cohen, 1995; Gopnik and Melzoff, 1997; Nichols and Stich, 2003), and some maintain that it may even be confined to human beings (Povinelli, 2000). For what it requires is a theory (either explicitly formulated, or implicit in the rules and inferential procedures of a domain-specific mental faculty) of the nature, genesis, and characteristic modes of causal interaction of the various different kinds of mental state. There is no reason at all to think that this theory should be easy to come by, evolutionarily speaking. And then on the assumption that the same or a similar theory is implicated in meta-cognition about one's own mental states, we surely shouldn't expect meta-cognitive processes to be very widely distributed in the animal kingdom. Nor should we expect to find meta-cognition in animals that are incapable of mind-reading." [185]

I've argued that intentional action essentially involves intentions, mental states which represent themselves as the cause of their objects. More strongly, my account of instrumental reasoning requires that agents have an idea of how these different mental states relate to each other. So I seem committed to the thoughts that a) all agents have a kind of metarepresentational state and b) all agents have the concept of a mental representation. Thus, if Carruthers is correct, then my theory of action is simply in conflict with the best understanding of animal action: animals act, but lack the conceptual sophistication my theory requires.

In particular, two kinds of experiments reveal the problems with my theory: the false belief task and the mirror test. The false belief task reveals that animals and small children lack the concept of belief. And the mirror test reveals that many animals lack self-consciousness. I respond: the empirical results have been misinterpreted. Neither the mirror test nor the false belief task has the purported import. The mirror test reveals that animals lack a certain class of self-locating representations, not that they lack them altogether. Similarly, the false belief task reveals that animals lack the full flowering of the theory of mind, not that they lack it altogether. Indeed, in order to make cognitive sense of the behavior of animals, we *must* postulate both self-locating representations, as well as partial theories of mind.

Before demonstrating this, I need to note one basic assumption I make (one which is in fact shared by Carruthers): animals have cognitions, i.e. represent the world and engage in reasoning. The challenge I am confronting is that animals genuinely act intentionally, but lack the representational states I postulate. If one denies that animals engage in representation and reasoning, i.e. cognition, then that would simply be grounds to deny that

[185] Carruthers 2008 p. 59

they act intentionally at all. If animals are just stimulus-response machines, the fact that they lack conceptual sophistication poses no challenge to my theory. So I here take it for granted that we are looking for cognitive explanations of the actions of animals. The question is *which* cognitive and representational capacities are needed to explain their actions.

5.3.1: The false belief task

Most adult human beings possess a relatively sophisticated capacity to attribute beliefs, intentions, and other mental states to other humans. Call this capacity a "theory of mind." The term is tendentious, insofar as it suggests that we do this via some either explicit or inarticulate theory, rather than perhaps some mechanism for simulation.[186] But set these questions aside for the moment: on any view, most adult humans can explain and predict the behavior of others on the basis of attributions of mental states – in particular, on the basis of *false* beliefs. I can say, "She thinks she has five dollars in her wallet, so she's buying some soba, but in fact she's forgotten that she spent it earlier on that knockoff poodle."

As it turns out, this capacity is shared by relatively few other animals, including human children under the age of four or so, as is revealed by the "false belief task."[187] In the "Sally-Anne" version of the task, children watch a play involving two dolls, Sally and Anne. Sally places a marble in a basket, and leaves; Anne moves the marble to a bucket. Sally returns, and the subject is asked: where will Sally look for her marble? If the subject answers "In the basket," then they pass the test; if they answer "In the bucket," then they fail. The results of this test are largely[188] insensitive to culture: human children the world over answer "in the bucket" until about age four. For obvious reasons, it is difficult to perform this particular experiment on animals. But it is possible to modify the setup somewhat: following a suggestion of Daniel Dennett's, researchers have constructed a situation where the subject animal needs to act based on a prediction about the behavior of a conspecific;[189] specifically, a situation involving competition for food. And chimpanzees, at least, fail this modified false belief task.[190] The traditional interpretation of these results is that children do not develop a theory of mind until around age four, when there is a

[186] cf. Carruthers 1996.
[187] In fact, there are a few variants on the task, e.g. the Sally-Anne task and the Smarties task.
[188] Though not completely. For instance, the deaf children of hearing parents show significant delays in theory-of-mind development (though the deaf children of deaf parents do not) (Wellman et al. 2011; Peterson et al. 2005).
[189] Dennett 1978.
[190] Kaminski et al. 2008.

sort of eureka moment and children realize "Aha! I can predict the behavior of others based on inner representational states, which may be true or false!" Animals, unfortunately, never quite manage that flash of conceptual insight.

As it turns out, the situation is somewhat more complex. Consider, for instance, that before children can pass the false belief task, they are capable of trying to deceive others. Indeed, recent research from Henry Wellman and David Liu suggests that, in the human case, there is a distinct developmental progression of theory-of-mind understanding. They write:

> "In particular, the studies confirm that theory-of-mind understandings represent an extended and progressive set of conceptual acquisitions. No single type of task - for example, false-belief tasks - can adequately capture this developmental progression."[191]

Instead, their research suggests that human children first understand that there can be a diversity of desires for a given object – i.e. that others act on the basis of their own desires, which may differ from the child's. They then understand that there can be a diversity of opinions on an unknown fact, and can predict action on the basis of ignorance.[192] They then begin to pass the false belief task. And, finally, they understand that a person's true emotions can differ from their outward emotions.

Similar complications apply to the animals. For instance, while chimpanzees fail the false belief task, they pass the knowledge-ignorance task: they can predict the actions of others based on the other's ignorance of some fact.[193] And a wide array of studies suggest that many other animals have some understanding of their own ignorance: they search out new information when they lack knowledge.[194]

What all this suggests is that while the full flowering of the theory of mind appears only in adult humans, a wide array of partial theories of mind appear in children and the other animals. This empirical point supports a more

---

[191] Wellman and Liu 2004 p. 537.

[192] As it turns out, the order of these two developments appears to be somewhat culture-bound. American, Australian, and German children first understand that there can be a diversity of opinions, and then latter understand ignorance. Chinese children proceed in the opposite order (Wellman et al. 2006).

[193] Kaminski et al. 2008.

[194] There are in fact several importantly different kinds of experimental design in this area. Some show that animals are capable of responding to uncertainty by choosing a sure but lesser reward when they are less confident about taking a test for a greater reward (e.g. Foote and Crystal 2007). Unfortunately, as Peter Carruthers suggests, we can give a first-order explanation of this behavior using the tools of rational choice theory and expected utility (2008 pp. 64-65). On the other hand, other studies involve the animal more or less asking for a hint when they are uncertain (e.g. Kornell et al. 2007). While we can also give an expected-utility explanation here, the explanation itself adverts to an understanding of knowledge and ignorance: the animal expects that, with new information, they will be able to answer correctly. That is, they can predict their own behavior in cases of knowledge and ignorance. In this case, Carruthers' alternative explanation is quite forced, and incapable of explaining why in some cases animals simply guess, and in others search out new information (2008 p. 66).

philosophical one: we can attribute thoughts involving a concept P to a creature that does not have a complete grasp of P. Indeed, we had better be able to, since it is unclear whether even adult humans have a *complete* grasp of their concepts! Consider, for instance, taking a square root. Many people who have passed through high school algebra have some grasp of the relation "is the square root of." But many of them also are only able to apply the concept in a few instances (say, 2 and 4), or only with the aid of a calculator, and even then only to certain limits - perhaps nobody is disposed to apply the concept correctly to extremely large numbers. More strongly, many square-rooters will be totally unfamiliar with *i* and the full (weird) flowering of "is the square root of" on the complex plane.[195] Yet, I suggest, none of this impugns the fact that the ordinary algebraist may know "Two is the square root of four" with perfect knowingness. Understanding a concept, being able to deploy it in thought, does not require a perfect or complete grasp of that concept: sometimes understanding is partial.

To return then to the topic of theory of mind in animals: perhaps it is true that such a theory achieves its full flowering in adult human beings, who are adept at deception, trickery, and the false belief task. But the fact that many animals have yet to master some of the Theory's more arcane components does not show that they lack it entirely. Rather, they have a partial grasp of the concept of mental state, and they clearly deploy this concept in predicting the behavior of other animals. So, I suggest, there is no obstacle to supposing that they deploy this concept when engaged in intentional action. That they deploy the concept in one stretch of thought does not entail that they are disposed to deploy it in every relevant stretch.

One might however respond: indeed, some animal behavior requires mental concepts as part of its cognitive explanation. But the representational theory is committed to a stronger thesis: that *all* action given a cognitive explanation involves such concepts. So merely discussing some particular kinds of behavior is not to the point. To respond to this, I want to consider a different sort of experiment: the mirror test.

5.3.2: The mirror test

The mirror test is quite simple: the subject is secretly marked with dye on a normally visible part of their body; how they respond to the dye establishes a baseline. They are then marked again, this time on a part of their body they cannot normally see. Then they are shown a mirror which reveals that they are so marked. If the subject then

[195] See Wilson 2006 Chapter 6 §VI, "Analytic Prolongation," for a more detailed discussion of the weirdness of $\sqrt{z}$. Wilson writes, "Due to Riemann is an evocative picture of the torsion that $\sqrt{z}$ evinces: imagine a ramped parking lot with two floors in which we can drive around forever without running into anything," (2006 p. 317).

responds to the new mark as they responded to the original, they pass the test; if not, then not. Adult humans and other great apes pass, as do bottlenose dolphins, orcas, elephants, and the European magpie; humans under 18 months fail, as does almost every other species of animal. The typical interpretation of these results is that creatures which fail the mirror test lack self-consciousness. Thus many have concluded that self-consciousness is in some way a cognitive achievement. If we imagine a scale of cognitive sophistication, purely world-directed, first-order thoughts lie at the bottom. Thoughts about oneself are located somewhere further up the ladder. What I wish to argue is this image is simply upside down: thoughts about oneself are the easiest and most natural of all; purely world-directed thoughts represent a cognitive achievement.

Consider this story from Ernst Mach's *The Analysis of Sensations*:

> "Not long ago, after a trying railway journey by night, when I was very tired, I got into an omnibus, just as another man appeared at the other end. "What a shabby schoolmaster that is, that has just entered," thought I. It was myself: opposite me being a large mirror. The physiognomy of my class, accordingly, was better known to me than my own."[196]

What is striking about this tale is that it essentially involves Mach failing the mirror test. What is perhaps stranger is that examples exactly analogous to this – Perry in the supermarket, Rudolf Lingens in the Stanford Library - are taken to be the proof that human beings have distinct first-person thoughts.[197] That is: when an animal fails the mirror test, we conclude that it must lack self-consciousness; when an adult human fails the mirror test, we conclude that it must *have* self-consciousness. This cries out for explanation. So I suggest: let's look at the details of each case more closely. What role in cognition and reasoning is played by first-person thoughts?

Imagine then that you are lost in a library.[198] Wandering the stacks, you come upon a map of the library. Unfortunately for you, a puckish prankster has removed the helpful "You are here" sticker. You have a perfect map of the library, and yet are still utterly lost. In this case, you have a wonderful set of world-directed beliefs about the library. You also have a wonderful set of demonstrative beliefs about where you are: "Those are some bookstacks over there," "There are no windows here," "I can walk over there if I want to," etc. What you are lacking is an indexical belief that would *connect* these two sets of representations, something to the effect of "This spot on the map is where I am."[199] In this situation, you are entirely capable of acting on the basis of your demonstrative beliefs, but utterly incapable of acting on the basis of your third-person, world-directed beliefs. This suggests a

[196] Mach 1897 p. 4. I have slightly emended the translation.
[197] This fact is discussed in detail by Perry 1979, among many others..
[198] This case obviously draws on Perry's example of Rudolf Lingens (1977 p. 492).
[199] This three-tiered structure draws on a suggestion of Perry's (1993 pp. 140-141). I don't wish to follow Perry's proposal in its full detail, but merely this basic insight.

basic reinterpretation of the mirror test: creatures that fail the mirror test don't lack all first-person beliefs; instead they lack those analogous to "This spot on the map is where I am," i.e. "That figure in the mirror is me." These are first-person thoughts which correlate the self with some object in a public world. That is: failing the mirror test reveals that the agent in question lacks *third person thought*.

To put the point another way: most writers fail to see just what a tremendous cognitive achievement a "purely world directed" thought is: it is a thought the content of which makes no mention at all of the agent, yet still manages to *be a thought with content*. So, for instance, "Rudolf Lingens is lost in the Stanford Library" is not a purely world directed thought: it does not specify *when* Lingens is there lost. We would need a way of specifying Lingens' location in objective space and time, without any reference to ourselves, before we arrive at a truly purely world-directed thought. A purely world directed thought, in Fregean terms, is this: a sense which would determine the same reference, no matter which agent in the world thought it. But it should become at once apparent that such thoughts tend to be hard to come by – the best candidates are beliefs about laws of nature (or about necessary truths).

If third-person thoughts are not directly relevant to action, why bother with them? Because, to paraphrase Perry, thoughts travel.[200] It is useful for me to be able to remember today what I did yesterday, to tell someone tomorrow what I did last week, and so forth. Or, to return to the earlier example: maps are really useful! They provide a general piece of information which can be brought to bear on a wide array of different situations; in this way, they are quite unlike indexical thoughts, which are in the limit one-use-only. This presents a rather different take on the fact that most creatures which pass the mirror test are social creatures. Socialiability is not a precondition on self-consciousness; rather, it is a precondition on other-consciousness. And this because social creatures have an interest in sharing thoughts with one another, so to speak. What I suggest therefore is that mirror failures lack *not* first-person thoughts in general, but rather a specific subset of them. They lack the ability to identify themselves as objects in a public world. But, as numerous cases from the literature on first-person thought show, such thoughts are not the essential part of first-person consciousness.

I've thus far argued that first-person thoughts are more basic than purely world-directed thoughts. But I haven't shown that metacognition is in some sense prior to purely world-directed thoughts. Perhaps animals are capable of having thoughts about themselves, but not about their thoughts. Here I respond: thought about thought is

[200] "Sentences, or tokens of them, travel," 1993 p. 140.

essential to action, in particular instrumental reasoning. Consider: the scrub-jay sees a delicious beetle. It flies over to the beetle, kills it, and eats it. We have here a structure of actions (flying over, killing, eating) organized around a goal (delicious-beetle-devouring). We might give a purely stimulus-response interpretation of this behavior: perception of beetle causes flying, perception of beetle nearby causes killing, perception of dead beetle causes devouring, etc. But, as I said earlier, this would rather undermine the challenge to my theory, since this thoughtless interpretation would also rather imply that the jay did not act intentionally. So let's consider a cognitive explanation of the jay's behavior. In particular: why doesn't the scrub jay, immediately on seeing the beetle, start chewing? Why isn't the order of action "eating, spearing, flying over"? There's an obvious answer, but an important one: because the jay knows it can't spear a beetle that it isn't standing next to. We might say: the jay knows "If I try to kill the beetle now, I'll fail," "If I fly over there, I'll be able to kill the beetle," etc. And these thoughts - about the relations among its actions, about the circumstances in which it will try and fail, etc. – these are thoughts not just about the jay as a whole, so to speak, but about particular tryings and how they would relate to one another.

In summary: failing the mirror test does reveal something important: the inability to bridge the gap between first-person and third-person thoughts. But such failure does not show that the subject has no first-person thoughts at all. Instead, first-person thoughts are necessary to make cognitive sense of any kind of intentional action.

5.4: FREUD'S INKWELL

> "It is quite obvious that grasping the wrong thing may serve a whole series of other obscure purposes. Here is a first example: it is very seldom that I break anything. I am not particularly dexterous, but by virtue of the anatomic integrity of my nervous and muscular apparatus, there are apparently no grounds in me for such awkward movements with undesirable results. I can recall no object in my home which I have ever broken. Owing to the narrowness of my study, it has often been necessary for me to work in the most uncomfortable position among my numerous antique clay and stone objects, of which I have a small collection. So much is this true that onlookers have expressed fear lest I topple down something and shatter it. But it never happened. Then, why did I brush to the floor the cover of my simple inkwell so that it broke into pieces?
>
> "My inkstand is made of a flat piece of marble which is hollowed out for the reception of the glass inkwell; the inkwell has a marble cover with a knob of the same stone. A circle of bronze statuettes with small terracotta figures is set behind this inkstand. I seated myself at the desk to write; I made a remarkably awkward outward movement with the hand holding the pen-holder, and so swept the cover of the inkstand, which already lay on the desk, to the floor.
>
> "It is not difficult to find the explanation. Some hours before, my sister had been in the room to look at some of my new acquisitions. She found them very pretty, and then remarked: 'Now the desk really looks

> very well, only the inkstand doesn't match. You must get a prettier one.' I accompanied my sister out and did not return for several hours. But then, as it seems, I performed the execution of the condemned inkstand."[201]

One need not be a Freudian to suspect that there are, at times, unconscious intentions - that we act on the basis of plans and goals of which we are not entirely aware. Theories of action like my own appear incompatible with this fact. I hold that every intention is a representational state, which represents itself as the ground of action: and how can such a state be unconscious? This is one reason to endorse a distinct-practical-attitude (DPA) theory of intention, which separates intention and representation of intention. On such a view, we therefore can say that an unconscious intention is simply an intention we do not represent.[202]

I respond: it would equally be a mistake to make intentions so distinct from representations that we could not see how we typically have accurate representations of them. On the DPA theory, there must be a story which says: in the normal case, we have accurate representations of our intentions. That is, in the normal case, agents infer (or quasi-infer) representations of intention from intentions. Unconscious intentions therefore are a breakdown of the normal inferential process: the intention is present, but some interfering factor prevents the agent from inferring a belief about that intention. However, once we put things this way, I can avail myself of basically the same account. DPA theories say that unconscious intentions are intentions which are inferentially isolated from the rest of one's beliefs. And I say the same: unconscious intentions are intentions for which the normal inferential processes have, for one reason or another, broken down.

One might rejoin: the normal inferential role is more essential to intentions as I've defined them than to distinct practical attitudes. That is, on the DPA theory, if these inferential processes break down, then an intention is still an intention. But on my view, intentions are in part constituted by their functional-inferential role in the agent's psychic economy. So it isn't possible that one should have a genuine representation which lacks these essential features: thus the representational theory does not really account for unconscious intentions. I respond: there is in fact no asymmetry here. For on either theory, it should be impossible for an agent's intentions to be normally unconscious: it would be utterly mysterious why such states would count as that agent's intentions, if in normal conditions she were systematically unaware of them. Thus even the DPA theory must, in some way, involve the

---

[201] Freud 1914 §VIII. David Velleman mentions the example, though the objection I am here responding to is not his (1996 p. 2).

[202] One might also interpret Freud's story as involving not an intention, but some other unconscious cause of behavior. While this might perhaps make problems for a view which, in some way, hold that all explanations of action are representational states, I do not endorse that view. On my view, there can be many causes of intentions which are not in turn instances of practical reasoning.

claim that one of the essential features of an intention is that, normally, the agent knows about it. And, on the my view, it is not as if an unconscious intention lacks *all* of its normal inferential properties. For instance, it might still play its familiar role in instrumental reasoning. Thus, on each theory, an unconscious intention involves both the breakdown and the normal functioning of largely similar inferential processes. So the my theory fares no better or worse than the DPA theory in this regard.

So much for unconscious intentions. But one might wish to raise a related objection: my view entails that if X believes "I intend to do A," then X intends to do A.[203] But this seems far too strong: surely an agent can be mistaken or self-deceived about their intentions. I respond: the entailment fails. On my view, an intention is not a person-centered representation such as "I intend to do A." Rather, it is a representation-centered representation such as "This very state will ground A." Thus, an agent might hold the former belief, but lack the latter representation. This is of course an inconsistent set of mental states, in the sense that it guarantees that one of the agent's representations is incorrect: but then, we wanted to say that self-deception is irrational; this is an explanation of why that should be so.

In summary: the view on which intentions are self-representations of causal efficacy can account for the ways in which intentions can, in one way or another, fail to be integrated into an agent's overall consciousness. In the case of unconscious intentions, the agent intends to do A, but this representation fails to figure in the normal set of inferences. In the case of self-deception, the agent does not intend to do A, but has a person-centered representation such as "I intend to do A." Since this representation is different from the representation I identify with intentions, my theory can account for this case as well.

5.5: NAIVE ACTION THEORY

The theory of action I've developed centrally involves intentions as mental states with a particular sort of content. This might be alleged to be incompatible with ordinary action-consciousness for two reasons. First, one might argue that it does not cohere with the implicit metaphysics of ordinary thought about action: a metaphysics of states is

[203] Bratman suggests an objection along these lines (2009 p. 515).

incompatible with a metaphysics of processes. Second, one might argue that it simply does not match the content of ordinary thought about action: no normal agent thinks in terms of self-referential representation states. I will address these objections in turn.

5.5.1: States and processes

Many contemporary philosophical accounts of intentional agency have recourse to intentions as mental states. But, Michael Thompson has argued, such "sophisticated" theories of action are fundamentally flawed. "Naive" action theory, centered on actions as worldly *processes* rather than mental *states*, is primary. Thompson, on my reading, presents two main arguments against sophisticated theories of action. First, they are not necessary: we can imagine a community of agents that only think in "naive" terms.[204] And second, they are not sufficient: sophisticated theories cannot account for the essentially aspectual character of ordinary action-talk.[205]

Thompson's case for the possibility of the naive community seems to rest on two ideas: first, that merely naive rationalizations serve as their instrumental reasoning, and second, that merely naive agents can, in effect, comprehend intentions for the future and intentions "in hiatus." Thus, naive agents come supplied with thoughts like "I am doing A because I'm doing B," "I did A because I was doing B," and so forth. Second, naive agents can think "I am doing A, just not right now," or "...just not yet," just as we sophisticates can say "The sycamore I planted ten years ago, it's growing well, it's overtaking the house," even in the middle of winter when the sycamore is not doing much growing at all.[206] This enables naive agents to register what we sophisticates would normally think of as mere intentions, e.g. "I am going to the opera tomorrow" as opposed to our "I intend to go to the opera tomorrow."

I don't wish to dispute these claims about the capacity of naive agents to represent instrumental connections and intentions for the future. Instead, I suggest that we need to explain not only what is present, but also what is not present. That is, even apparently naive agents are not forever flailing around, taking terrible means to their ends; they don't try to leave the room by simply walking through the doorway, but open the door first. In order to explain this obvious fact, I suggest that agents also need to have thoughts about what they do not do; in the above example, the agent knows that if they try to walk through the doorway without opening the door, they will not

[204] 2008 II.8.6.
[205] 2008 p. 133.
[206] 2008 p. 141.

get out of the room. And this sort of knowledge is not naive knowledge: it is about the causal connection between a mental state (intending to walk through a closed door) and the world. So, I conclude: even naive agents and ordinary human beings have a wide variety of sophisticated thoughts.

Thompson's second argument is that sophisticated theories cannot account for aspect. This merits some explanation. In brief, Thompson distinguishes between states and processes. States, e.g. "is red," admit of distinctions of tense: the apple is red, the apple was green. Processes, e.g. "eating," admit of distinctions of tense and aspect: the cat is eating the mouse, the cat was eating the mouse, the cat ate the mouse. I take it for granted that Thompson is right about this: everyday action-talk involves aspects and processes, not just states, and this talk is not reducible to talk about states. What I wish to show is that this is compatible with a sophisticated theory of action, on which the mental state of intention is primary.

The basic idea is this: intentions are states which are the cause (or ground) of processes.[207] Consider the guidance of action by instrumental reasoning. Intentions guide action by providing for the standing possibility of instrumental reasoning - of taking further actions in pursuit of one's goals, as the agent takes to be appropriate at the time. Thus, an intention directed at some general end – getting to the store – issues (if the agent is rational and effective) in further actions suited to that end – putting on shoes, opening the door, walking down Negley Avenue, and so forth. The process Thompson sees here is just the process of instrumental reasoning. And the existence of such a process is of course compatible with its failure to terminate in the intended end. On this view, then, Thompson's three judgments (X is doing A, X was doing A, X did A) all share a reference to the mental state of intention. The differences mark out differences in the causal efficacy of that state. Thus, intentional action is not merely a state, nor merely a process, but involves a complicated interaction between mental states and the processes which they govern.

Thus, I take it that the sophisticated theory of action, at least as I have developed it, is compatible with Thompson's claims about aspect and processes. While traditional sophisticated theories, as exemplified by Davidson, may have focused on eye-blink basic actions, and may have mistakenly suggested that intentions are propositional attitudes, the theory I have here developed is committed neither to the possibility nor the ubiquity of

[207] There are perhaps some difficult metaphysical issues lurking here. Thompson remarks on "causality appropriate to states" (2008 p. 133); if states and processes necessarily involve different kinds of causality, then that might pose a problem for my account. However, I differ from Thompson on this score: there is no such thing as "state-causation" as distinct from "process-causation." One of the great virtues of the Woodwardian account of causation developed in Chapter One (§1.2) is that it prescinds from any metaphysical or ontological questions about causal relata: it merely presupposes that the values of these variables can be related in certain specific ways.

basic actions, nor does it hold that the content of an intention is some proposition. So while the theory is undeniably sophisticated insofar as it centers on the mental state of intention, it hopefully does not suffer from all the same flaws as prior such theories.

5.5.2: Ordinary practical thought

Inspired by Thompson's work, one might form a much less metaphysical objection to my theory of action: it simply does not cohere with ordinary thought about action. I talk of intentions as mental states which refer to themselves, a conception of dependence derived from the idea of a causal graph, the conditional causal efficacy of instrumental beliefs, and a variety of other jargon-laden technical concepts. But no ordinary agent thinks in these terms!

Of course, I accept that nobody – not even me – apprehends action in these terms. So I must answer the question: how does the theory of action here developed relate to ordinary thought about action? Here is one answer I reject: the way that classical mechanics relates to folk physics. If folk physics is a theory, then it has the same subject matter as classical mechanics – matter in motion, we'll say. And classic mechanics supersedes folk physics insofar as it is a more accurate representation of that underlying material. But this is not the case with action. Ordinary thought about action is not a theory of some possibly unknown topic, "intentional action." Rather, ordinary practical thought is itself part of the subject matter of any possible theory of intentional action. So a philosophy of action cannot hope to simply supersede and replace ordinary practical thought. But by the same token, neither can it be entirely disjoint. Philosophy of action cannot be a theory of something other than ordinary practical thought, since that constitutes an essential part of intentional action.

Thus, I say: philosophical action theory aims to make explicit, to give a perspicuous philosophical account of, material which is already implicit in ordinary practical thought. My starting points are simple, familiar. The many say: there is a difference between intended and side effects. I add: here is an account of this distinction, which can be used to decide hard cases. The many say: you should take the means to your ends. I add: yes, you should - given these qualifications. The many say: typically, we know what we're doing. I add: here's how. The technical terms and arcane jargon I rely on only exist to make plain what is already present in ordinary practical consciousness. The arguments of the preceding chapters have all been directed to that end: their point is that other, apparently simpler accounts of practical categories, or instrumental reason, or practical knowledge, simply fail to correctly capture the phenomena which every agent already apprehends.

5.6: SUMMARY AND CONCLUSION

In this chapter, I've responded to a number of objections to the account of intentional action I've developed. Most of them could equally be leveled at any similar theory of action. However, in responding to them, I've relied on the details of my account. Hopefully, this provides some further confirmation of the overall account.

6.0: PUTTING IT ALL TOGETHER

6.1: INTENTIONAL ACTION

Up to now, I've presented three largely independent arguments about three aspects of intentional action: practical categories, instrumental reason, and practical knowledge. In each case I argued that thoughts which represent themselves as the grounds of action are both necessary and sufficient to account for the phenomena. The account of intentional action I now forward simply generalizes these ideas. I say: practical thought represents itself as the ground of its object. This basic idea is both necessary and sufficient to explain intentional action. Practical thought includes intentions, which represent themselves as the cause of some worldly goings-on, and instrumental beliefs, which represent themselves as the cause of intentions, on the condition that those intentions are instrumentally rational. Taken together, these representations ground processes of instrumental reasoning; this is the characteristic way in which agents guide the world towards desired outcomes. If such representations are formed on the basis of sound reasoning, which includes inter alia a rational expectation that they will be true once formed, then they are knowledge. Thus, practical thought is both a knowledgeable representation of, and a ground of, world-goings on: intentional actions. That is the schematic outline of the account.

However, it might be challenged. In particular, no argument thus far has shown that there is not some further essential feature of intentional actions. One might accept all the arguments in the foregoing chapters, but wish to add something more to intentional agency. Consider what I called in Chapter One (§1.1) "rational autonomy," which I glossed as the distinctly human exercise of free reason - perhaps the capacity to respond to the reasons that there are (hereafter Reasons), or perhaps to have beliefs about such Reasons, or perhaps something else. Many authors have argued that rational autonomy is also an essential feature of intentional agency; if this is so, the account I sketched above is false, or at least incomplete.[208]

[208] There are of course other potential candidates for a fourth feature of intentional action; while I focus on rational autonomy, the arguments that follow apply mutatis mutandis to many other such candidates.

The argument that my account is correct, and that rational autonomy is not an essential feature of intentional action, begins from the premise, mentioned briefly in Chapter One (§1.1), that intentional action is not a chimera. Its essential features are not united by mere conceptual fiat. If something is a green book, then it necessarily follows that it is green and a book, but these features have little to do with another: a green book is just a book which happens to be green. Intentional action, by contrast, is not merely a stretch of instrumental reasoning which happens to involve practical knowledge and practical categories. Its essential features come as a package.

The argument then runs as follows: practical categories, instrumental reason, and practical knowledge are essential and inseparable features of intentional action. The proper account of these features is that practical thought represents itself as the ground of its object. But, if neither the capacity to form representations nor the concept of ground involves rational autonomy, then rational autonomy is separable from the practical categories, instrumental reason, and practical knowledge. So rational autonomy is not an essential feature of intentional action. So the practical categories, instrumental reason, and practical knowledge are the only essential features of intentional action. So the full account of intentional action is: practical thought represents itself as the ground of its object. There are two places one might wish to object: first, to the claim that rational autonomy is separable from my three preferred features of intentional action, and second to the claim that these three features really are inseparable. I'll respond to these objections in turn.

As I suggested before, one might argue that rational autonomy is already implicated in the capacity to form representations at all, or in the concept of dependence. The latter claim is easy to reject: the basically Woodwardian conception of dependence I rely on (discussed in Chapter One (§1.2)) clearly does not rely on Reasons, nor does any leading theory of causation do so. On the other hand, many have suspected that beliefs themselves already involve some operation of Reason. While this issue is somewhat beyond the scope of this dissertation, I suggest that the model of practical reasoning I develop here can be extended to account for theoretical reasoning as well – for perception and inference, belief and representation in general (I speculate on this possibility in slightly more detail in §6.4). In short: if we didn't need to appeal to Reason to account for the distinctive features of practical thought, then I suggest that we can do the same for theoretical thought.

There is one very straightforward argument for the claim that these three features are inseparable: since the account of each is the same, as argued for in Chapters Two through Four, then they are inseparable. But even prescinding from the details of my account one can see their interconnections. Instrumental reasoning and practical

categories are clearly deeply intertwined. The characteristic import of such categories concerns instrumental reasoning; for instance, one reasons from intentions, but not foreseen side effects. When things do not go according to plan, that is a mistake, and it means I need to develop a new plan; when doing so, I do not aim to maintain the side effects of my earlier plan, but merely whatever the ultimate goal was. And practical knowledge and instrumental reasoning are also closely connected. First, instrumental reasoning seems essential to the possibility of practical *knowledge*. Like most knowledge, the epistemic credentials of such knowledge depend on how it is formed. So we have practical knowledge when we form intentions on the basis of sound instrumental reasoning: if I form the intention to do A, knowing that I will bring A off if I intend to, then I know: "I am doing A." Second, instrumental reasoning seems to require practical knowledge: if I cannot know the likely effects of my intentions, or whether I am now doing A, it would be impossible to effectively take the means to my ends. So, I conclude: these three features of intentional action are essentially intertwined.

So I conclude: practical thought represents itself as the ground of its object.

6.2: METAETHICS

In the previous section, I argued that intentional agency is a relatively minimal notion. Rational autonomy, the capacity to respond to or have beliefs about Reasons, is at best a particular kind of intentional agency. Does this have metaethical implications? According to Kieran Setiya, it does: he takes a very similar account of intentional action to refute constitutivist programs in metaethics, such as Christine Korsgaard's neo-Kantian project. According to Setiya, constitutivist programs hope to derive substantial moral norms from the nature of intentional agency. But since intentional agency is such a minimal notion, no such norms are forthcoming.

While I share Setiya's premise (I here ignore the slight differences between our accounts of intentional agency), I dispute his conclusion. This disagreement stems from our different readings of constitutivist ambitions. On Setiya's reading, Korsgaard aims to derive moral standards from the nature of mere agency. Pace Setiya, I suggest that Korsgaard (and others) aimed to derive moral standards from a the nature of specifically *human, reflective* agency in particular, not intentional agency in general. Consider Korsgaard's discussion of animal agency in *The Sources of Normativity* and *Self-Constitution*. Here is a representative passage:

> "The human mind *is* self conscious-in the sense that it is essentially reflective. ... A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in

> conscious activities, but it is not conscious *of* them. ... But we human animals turn our attention on to our perceptions and desires themselves, on to our own mental activities, and we are conscious *of* them. That is why we can think *about* them. And this sets us a problem no other animal has. It is the problem of the normative."[209]

Thus, on Korsgaard's view, animals act, are agents, but are not subject to moral norms. This strongly suggests that Korsgaard never aspired to derive moral norms from *mere* agency, but relied on some more demanding notion - reflective agency. So I do not think that the constitutivist project is immediately refuted by the account of intentional action developed here.

One might object that there has been some terminological slippage here, from "intentional agency" to "mere agency."[210] While *mere* agency includes animals, *intentional* agency is the specifically human phenomenon. On this view, an analysis of intentional action already is an analysis of specifically human agency. So the fact that, on my account, intentional action is such a minimal notion shows that human agency is likewise too minimal to ground moral norms.

I respond: this is not how I meant to use the notion of intentional agency. As discussed in Chapter Five (§5.3), I think the account here applies to the actions of animals and small children. It was never my intention to analyze reflective human agency; my aim was to analyze intentional agency in general. So the account *here* developed is not an account of human agency, and it leaves open the possibility that there is a particular kind of intentional agent, one characterized by rational autonomy, which is by its nature subject to some substantial moral norms. However, one might have some *further argument* which shows a) there is no such notion as rational autonomy b) human beings are not autonomous or c) the notion of rational autonomy does not ground substantive moral norms. Nothing I have said here supports or undermines any of (a-c). So while an argument for (a-c) would undermine the constitutivist project, nothing about the account of intentional action here developed does.

However, this account does suggest a particular burden for constitutivist theories. If moral standards were derivable from mere agency, that would be an extraordinarily strong foundation for morality: there would just be no escaping it. If however there are other kinds of agency, ones which are not subject to moral norms, then it seems possible that you or I might become such unreflective agents. If moral standards are to be self-supporting, then they must give reflective agents *reason to be reflective*.[211] We can see the force of this problem by considering, briefly,

---

[209] Korsgaard 1996 pp. 92-93.
[210] Thanks to Kieran Setiya for this objection.
[211] I should note that some constitituvists, such as Korsgaard, take up this challenge; on my reading, this is the point of her obscure discussion of suicide in Lecture 4 of her (1996).

the Pyrrhonist skepticism of Sextus Empiricus. Sextus argues, in effect, that being reflective is by its own lights *bad for us*. Careful consideration of skeptical arguments lead to suspension of judgment on such matters, and thus to a life without Reasons. According to Sextus, such a life, free from the anxiety induced by reflection, is superior to the reflective life.[212] If they wish to avoid this conclusion, constitutivists must show not only that rational autonomy grounds substantial moral norms, but that these norms give reflective agents to go on being reflective. Unfortunately, I don't here have space to discuss the prospects of discharging this burden; that must await future work.

And this account of mere intentional action also has implications for our account of distinctively human agency. Consider again the above quote from Korsgaard. In it, she emphasizes that human beings are self-conscious, i.e. can have thoughts about their own mental states. On her view, this is the distinctive feature of human agents which grounds moral obligation. But, on the account I've developed, that cannot be right. Agents in general are self-conscious, in the minimal sense of having representations of their own representation states: this is an essential part of instrumental reasoning. As I discuss in Chapter Five (§5.3), we must impute to the parrot an understanding of how its thought effects the world in order to make sense of the orderly and structured progression of its intentional actions. So self-consciousness, in this sense, is not a distinguishing feature of human agency. This suggests that what Korsgaard calls "reflectivity" is not adequately capture by the notion of self-consciousness.[213] Again, I do not take this point to refute constitutivist theories, but merely to show the shape of the challenges they face.

I've thus far focused on constitutivist accounts of moral norms; I should also note that the account of instrumental beliefs given in Chapter Three suggests the possibility for a constitutivist account of instrumental norms. But I am yet thoroughly uncertain about how such an account might proceed, so I won't speculate further here. In summary: while this account of intentional agency does not immediately refute constitutivist ambitions, it does suggest two important challenges that such accounts must meet.

---

[212] Sextus Empiricus 2000 Book I §XII.

[213] Sometimes Korsgaard herself seems to suggest as much – she remarks that some animals may be conscious of themselves, but that her idea of self-conscious is more closely tied to the ability to ask normative questions about one's perceptions and desires (Korsgaard 2009 §§6.1.5-8)

6.3: ETHICS

So much for the metaethical implications of this account. Does it have first-order ethical implications? Yes; I'll here consider just two, concerning the much-contested doctrine of double effect and the distinction between doing and allowing. Both of these ideas have been subject to serious criticism by consequentialist thinkers. Sometimes the objection is straightforward: these distinctions just don't matter, and we should ignore them. But sometimes the objection is more difficult: these distinctions are incoherent, and so we should ignore them. There is no philosophically rigorous account which could support the first-order work these distinctions have been supposed to do – or so this objection runs. It is this latter issue I here address.

6.3.1: The doctrine of double effect

Consider the doctrine of double effect first.[214] In the abstract, this principle states that there are some outcomes which it is permissible to foresee as side effects of pursuing some other permissible goal, but which it is not permissible to intend as means to that goal. Perhaps the clearest application of the doctrine is the distinction between the Strategic Bomber and the Terror Bomber, two men winging their nighttime way towards Occupied Europe. Strategic Bomber has been sent to destroy a munitions factory; he knows that this factory is surrounded by a civilian neighborhood, and his bombs will result in many innocent deaths: but destroying a munitions factory in the prosecution of a just war is permissible, despite the tragic side effects. Terror Bomber, by contrast, has been sent to destroy a civilian neighborhood, in the hopes of reducing enemy morale and ending the war. He knows of course that there is a nearby munitions plant which will be destroyed by his bombs: but his target is the innocent lives living around it. According to the doctrine of double effect, Strategic Bomber's action is permissible, while Terror Bomber's is forbidden, despite the fact that both acts have the same consequences.

That, in outline, is the doctrine of double effect. I here defend it, in this sense: there is a rigorous distinction between Strategic Bomber and Terror Bomber. In Chapter Two, I developed an account of the distinction between intended and side effects. An intended effect is just some effect E produced by a representational state of the form "This very representation grounds E," i.e. an intention. Side effects are those effects of intentions which are not intended. We can apply this analysis to cases of instrumental reasoning, i.e. structured sets of intentions and effects. The agent recognizes some chain of dependence running from their

[214] I discuss the doctrine in more detail in my "A Disreputable Doctrine," unpublished.

intention for the end, through the means, and to the end; the existence of this chain is grounded in the very recognition of the existence of that chain. Or, to follow the above form, "This very representation grounds E via such-and-such a chain of dependence." Any effects lying on that directed path from the agent's intention to their end are intended; effects of their practical thought which are not intended are side effects. This allows that there may be other chains of dependence running from their intention to their end which are not intended: the agent recognizes such chains, but their recognition of such chains are not part of the explanation of the existence of such chains. Put another way, their recognition of such chains do not play a role in their practical reasoning.

This, I take it, establishes a philosophically rigorous distinction between intended and side effects, a distinction which could support the doctrine of double effect. But there remain a number of objections. A first objection, easily dismissed, is epistemic: it can be difficult to discern the precise content of an agent's intentions. But there are epistemic difficulties just about everywhere, and in many cases we can tell quite well what a person's intentions are.[215] So I do not take this to be a serious problem.

A second objection runs as follows: Terror Bomber might say in his defense "It was not my intention to kill the civilians, but merely to make them *appear* dead, since that was all that was necessary for my purposes. Unfortunately as a side effect of making them appear dead they actually did die - but if they somehow miraculously survived while merely appearing dead, that would have been fine with me!"[216] I respond: this again proves the usefulness of the Woodwardian notion of causation. Given the *actual causal* circumstances of Terror Bomber's actions, an intervention on the deaths would have wrecked his plans (and he knows this), and this is true despite there being other possible worlds in which it would not have. What we look to in determining intended and side effects are the *actual* dependence relations in the agent's world, not relations in other possible circumstances.

A third objection holds that the definition I've given of side effects is no good, since it fails to capture the following case: a group of spelunkers are trapped in a cave. In this case, one of the cavers is stuck in a narrow passage; water is rising and the only way for the five behind him to get him out of the passage (and thus avoid death) is to chop him up (whereupon he will shortly die of blood loss). Yet on my definition of side effects, the killing of

[215] As Anscombe points out, these differences typically manifest in different patterns of practical reasoning (1957 §25). While there are perhaps cases where different underlying intentions produce the same surface pattern, they are distinguished by their counterfactual implications, and again, while counterfactuals can be difficult to determine, they are not generally impossible.
[216] Bennett 2001 p. 101.

the man is a side effect of their actions, since his actual death is causally irrelevant to their ends. And one might find this conclusion mad – surely these cavers may not chop up their unwilling partner to save their own skins!

I respond: by my criterion, the killing of the man is a side effect of chopping him up. This does not establish that it is permissible to chop him up. Consider a different case: X sneaks into Y's room at night and secretly eats Y's arm. Before morning comes and Y wakes, X replaces the original arm with a perfectly functional duplicate and leaves, Y none the wiser. I suggest that a defender of double effect can plausibly say: X's actions wildly violate Y's rights to bodily integrity. Similarly, the stuck man has both the right to life *and* the right to bodily integrity. Consequently, chopping him up – which is intentional, and not a side effect – violates his rights and is therefore wrong. So the doctrine can explain why this action is wrong, despite his death being a side effect.[217]

In other words, the doctrine of double effect, coupled with the moral view that the only wrong action is the intentional killing of the innocent, is plainly inadequate. But the implausibility here lies with such a restricted moral view, not the doctrine. So just because some action does not include intentional killing does not imply that it is perfectly fine; it may be wrong because it is wrong in itself, or it fails to take sufficient account of alternatives, or the good end is not proportionate to the bad side effect, or for perhaps other reason.

I have not here canvassed every objection to the doctrine of double effect; that must await future work. But I hope to have shown how the conception of side effects developed here can help answer some traditional objections to the doctrine.

6.3.2: Doing and allowing

Consider next the distinction between doing and allowing (e.g. between killing a person and letting them die). This distinction, I suggest, is not grounded in the nature of action. This is not to say that there is no difference between doing and allowing; perhaps there is even a philosophically rigorous one. But since practical reason treats these two classes of action identically, and if ethics is in the business of evaluating action and practical reasoning, then it's not clear what grounds there could be for a general moral difference between doing and allowing.

[217] Of course, one might also hold that the action is *not* wrong, since in this instance chopping him up does not amount to an *unjust* injury (cf. Thomson 1971. See also my "Rights, Wrongs, and Crimes against Humanity" (unpublished) for a more detailed look at related cases).

Consider the following cases:[218]

> *Rockslide One*: H has reached the top of a scree slope before her companion UM. She notices a boulder begin to roll down the scree; she knows if she does not reach out and stop it, it will precipitate a rockslide which will kill UM. Because she wishes to inherit UM's vast fortune, she does not stop the rock: and everything transpires as expected.
>
> *Rockslide Two*: H has reached the top of a scree slope before her companion UM. She notices a boulder begin to roll towards her; she knows if she gets out of the way, it will precipitate a rockslide which will kill UM (whereas if she stands still nobody will be hurt). Because she wishes to inherit UM's vast fortune, she does not stop the rock: and everything transpires as expected.
>
> *Rockslide Three*: H has reached the top of a scree slope before her companion UM. She notices a boulder which, if pushed, will precipitate a rockslide which will kill UM. Because she wishes to inherit UM's vast fortune, she pushes the boulder: and everything transpires as expected.

It seems clear that H kills UM in *Rockslide Three*. It seems equally clear that she lets him die in *Rockslide One*. And *Rockslide Two* is difficult to decide. But in each case, I suggest, the practical reasoning of H is identical. She has an end: inherit a vast fortune. In each case, she recognizes a causal pathway from her practical reasoning to that end. The particulars differ (don't stop the rock, get out of the way, push the rock), but the *causal structure* of each situation is the same. This follows from the basic Woodwardian conception of causation in play here, which does not distinguish between "causation by omission" and "causation by comission."[219] Despite the fact that there is no e.g. transfer of energy or momentum from H to the rockslide in the first two cases, her practical reasoning is a cause of the rockslide, insofar as had she reasoned differently, the rockslide would not have occurred. In every case, H is a self-conscious cause of the rockslide and the death.

Thus I conclude: the distinction between doing and allowing is not grounded in the concepts of mere intentional agency. Practical reason treats doings and allowings identically. So there is little prospect for a serious moral distinction between these two classes of action.

---

[218] These cases are obviously adapted from James Rachels' (1975). Rachels' argument has been the target of much criticism; I suspect this is because he moves directly from his intuition that the agent in each case is equally morally blameworthy to the conclusion that there is no distinction between doing and allowing. I make a different argument: because the practical reasoning of each agent is identical, as shown by a prior theory of action, our moral evaluation of each agent should be identical. This does not show that there is no distinction between doing and allowing, but merely that it does not make a moral difference.

[219] cf. Woodward 2003 Chapter 2 §8 for an extensive discussion of these issues.

6.4: SPECULATIONS

I've thus far discussed the general account of intentional action I've developed over the course of this dissertation, as well as some of its implications for ethics and metaethics. But a general account of intentional action is still only a partial account of the mind of a rational agent. So I want to conclude by suggesting some ways such an extension might go. I'll consider three other aspects of the mind: perception, inference, and self-knowledge.

Consider first perceptual judgment. I look out my window and see the fireworks over Shadyside. I arrive at the knowledge that they're shooting fireworks up from the Shadyside Nursery on the basis of experience. Yet giving a philosophical account of what I just described has proven difficult. Consider, for instance, the notorious Gettier Problem. In essence, Gettier noticed that it was possible to arrive at a true belief, and do so according to ordinary standards of justification, yet for the relation between the belief and its object to be somehow accidental. The difficulty has been spelling out what in the world it is to be nonaccidentally in touch with the world. On its face, then, this problem in the theory of perceptual judgment shares some features with the problem of deviant causal chains, of which I proposed an account in Chapter Two. I suggest that the analogy between these two problems is promising: they are both cases where we want to specify "the right kind" of dependence between thought and world. So if the right kind of practical dependence involves the agent's knowledge of that dependence, then perhaps the right kind of perceptual dependence also involves the agent's causal understanding of how their perceptions relate to their objects. Or, so I hope to argue in future work.

Consider next the topic of inference. Oftentimes we form beliefs not about objects of perception, but about objects which go beyond perception – objects and events which are too small, or too far in the past, or which do not yet exist. And we do this by inference from things known by other means. This topic – reasoning from belief to belief – has also proved somewhat vexed. What are the rational standards of coherence? And what can ground them? What I suggest is that it is possible to generalize the account of instrumental beliefs I developed in Chapter Three to cover inference in general. But a detailed investigation of this possibility must await future work.

Last, consider the topic of self-knowledge. It has long been thought that we know our own mental states – beliefs, desires, imaginings, emotions and so forth – in a way which is different from the way we know the world. But accounting for this kind of self-knowledge has also proven difficult. In Chapter Four, I proposed an account of self-knowledge of action and intention, which allowed that there is a way of knowing our intentions which is also a

way of knowing our actions. What I hope is that this account can be extended to belief generally: that it can be shown that beliefs (like intentions) are, of themselves, ways of knowing about both ourselves and the world.

Of course all these ideas are quite speculative. I merely wish to suggest that the account here given of intentional action shows promise in shedding light on other vexed questions in the philosophy of mind generally.

6.5: THAT'S ALL, FOLKS

I began with the promise, and the problem, of action – the promise, that a turn to action would resolve some of my own longstanding worries about morality and its foundations – the problem, that the conceptual landscape of the philosophy of action was deeply confusing. After much tribulation I arrived at an account of action, and an idea of how that confusing conceptual landscape in fact amounted to a number of partial views of the same phenomenon. Yet this ending is bittersweet, since I feel I've understood action only to discover that its promise was rather an illusion. While I take the account I've developed to clarify the possibilities for a constitutivist foundation of moral obligation, it hardly determines whether such a possibility is actual. But perhaps I should not be surprised that there is still more philosophy to do.

BIBLIOGRAPHY

Anscombe, Elizabeth; 1957 (2000). *Intention*. Harvard: Harvard University Press.

Anscombe, Elizabeth; 1961. "War and Murder," pp. 44-52 in *Nuclear Weapons: A Catholic Response* ed. Walter Stein. London: Sheed and Ward.

Anscombe, Elizabeth; 1969. "On Promising and Its Justice, and Whether It Needs be Respected In Foro Interno," pp. 61-83 in *Crítica* vol. 3 no. 7/8.

Anscombe, Elizabeth; 1971. *Causality and Determination: An inaugural lecture*. CUP Archive.

Anscombe, Elizabeth; 1975. "The First Person," pp. 45-65 in *Mind and Language* ed. Samuel Guttenplan. Oxford: Clarendon Press.

Anscombe, Elizabeth; 1978. "Rights, Rules, and Promises," pp. 318-323 in *Midwest Studies in Philosophy* vol. 3.

Anscombe, Elizabeth; 1983 (2005). "The Causation of Action," pp. 89-108 in *Human Life, Action and Ethics*, ed. Mary Geach and Luke Gormally. Imprint Academic.

Arpaly, Nomy; 2006. *Merit, Meaning, and Human Bondage*. Princeton: Princeton University Press.

Bennett, Jonathan; 2001. "Foreseen Side Effects versus Intended Consequences," pp. 85-118 in *The Doctrine of Double Effect*, ed. P.A. Woodward. Notre Dame: University of Notre Dame Press.

Bishop, John; 1989. *Natural Agency*. Cambridge: Cambridge University Press.

Boyle, Matthew; 2011. "Transparent Self-Knowledge," pp. 223-241 in *Proceedings of the Aristotelian Society Supplementary Volume* LXXXV.

Brand, Myles; 1984. *Intending and Acting*. Cambridge: The MIT Press.

Bratman, Michael; 1987 (1999). *Intention, Plans, and Practical Reason*. CSLI Publications.

Bratman, Michael; 1991. "Cognitivism about Practical Reason," pp. 117-128 in *Ethics* vol. 102 no. 1.

Bratman, Michael; 2009. "Setiya on Intention, Rationality, and Reasons," pp. 510-521 in *Analysis Reviews* vol. 69 no. 3.

Broome, John; 1999. "Normative Requirements," pp. 398-419 in *Ratio* XII 4.

Broome, John; 2007. "Wide or Narrow Scope?" pp. 359-370 in *Mind* vol. 116 no. 462.

Broome, John; 2013. *Rationality through Reasoning*. Malden: Wiley-Blackwell.

Buss, Sarah; 1999. "What Practical Reasoning Must Be if We Are to Act for Our Own Reasons," pp. 399-421 in *Australasian Journal of Philosophy* vol. 77 no. 4.

Carruthers, Peter; 1996. "Simulation and Self-Knowledge," pp. 22-38 in *Theories of Theories of Mind*, ed. Peter Carruthers and P.K. Smith. Cambridge: Cambridge University Press.

Carruthers, Peter; 2008. "Metacognition in Animals," pp. 58-89 in *Mind and Language* vol. 23 no. 1.

Castañeda, Hector-Neri; 1966. "He: A Study in the Logic of Self-Consciousness," pp. 130-157 in *Ratio* vol. 7 no. 2.

Cokely, Edward; Feltz, Adam; 2009. "Individual Differences, Judgment Biases, and Theory-of-Mind," pp. 18-24 in *Journal of Research in Personality* vol. 43 no. 1.

Davidson, Donald; 1963 (2001). "Actions, Reasons, and Causes," pp. 3-20 in *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, Donald; 1969 (2001). "How is Weakness of the Will Possible?" pp. 21-42 in *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, Donald; 1971 (2001). "Agency," pp. 43-62 in *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, Donald; 1973 (2001). "Freedom to Act," pp. 63-82 in *Essays on Actions and Events*. Oxford: Oxford University Press.

Davidson, Donald; 1978 (2001). "Intending," pp. 83-102 in *Essays on Actions and Events*. Oxford: Oxford University Press.

Dennett, Daniel; 1978. "Beliefs about Beliefs," pp. 568-570 in *Behavioral and Brain Sciences* vol. 1.

Dennett, Daniel; 1996. *Elbow Room: The varieties of free will worth wanting.* Cambridge, MA: MIT Press.

Eells, Ellery; 1982. *Rational Decision and Causality*. Cambridge: Cambridge University Press.

Egan, Andy; Elga, Adam; 2005. "I Can't Believe I'm Stupid," pp. 77-93 in *Philosophical Perspectives* no. 19.

Falvey, Kevin. "Knowledge in Intention," pp. 21-44 in *Philosophical Studies* vol. 99 no. 1.

Falvey, Kevin. "The Cause of What It Understands," unpublished.

Fine, Kit; 1994. "Essence and Modality," pp. 1-16 in *Philosophical Perspectives* vol. 8.

Finlay, Stephen; 2010. "Against All Reason?" pp. 155-178 in *Hume on Motivation and Virtue*, ed. Charles Pigden. Palgrave MacMillan.

Foote, Allison; Crystal, Jonathon; 2007. "Metacognition in the Rat," pp. 551-555 in *Current Biology* vol. 17.

Ford, Anton. "Varieties of Naive Action Explanation." Unpublished.

Frankfurt, Harry; 1978. "The Problem of Action," pp. 157-162 in *American Philosophical Quarterly* vol. 15 no. 2.

Freud, Sigmund; 1914. *The Psychopathology of Everyday Life*. trans. A.A. Brill. New York: The MacMillan Company.

Garson, James, "Connectionism", *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), ed. Edward N. Zalta, forthcoming. URL = <http://plato.stanford.edu/archives/spr2015/entries/connectionism/>.

Gopnik, Alison; Wellman, Henry; 2012. "Reconstructing Constructivism," pp. 1085-1108 in *Psychological Bulletin* vol. 138 no. 6.

Grice, H.P.; 1971. "Intention and Uncertainty," pp. 3-19 in *Proceedings of the British Academy* vol. LVII.

Gugliemo, Steve; Malle, Bertram; 2010. "Can Unintended Side Effects Be Intentional?" pp. 1635-1647 in *Personality and Social Psychology Bulletin* vol. 36 no. 12.

Hancox, J.S. "A Disreputable Doctrine," unpublished.

Hancox, J.S. "Rights, Wrongs, and Crimes against Humanity," unpublished.

Harman, Gilbert; 1976. "Practical Reasoning," pp. 431-463 in *The Review of Metaphysics* vol. 29 no. 3.

Harman, Gilbert; 1986. "Willing and Intending," pp. 363-380 in *Philosophical Grounds of Rationality*, ed. Richard Grandy and Richard Warner. Oxford: Clarendon Press.

Harper, William; 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory," pp. 25-36 in *Erkenntnis* vol. 24 no. 1.

Heckhausan, Heinz; Beckman, Jürgen; 1990. "Intentional Action and Action Slips," pp. 36-48 in *Psychological Review* no. 97.

Hempel, Carl; and Oppenheim, Paul; 1948. "Studies in the Logic of Explanation," pp. 135-175 in *Philosophy of Science* vol. 15 no. 2.

Hempel, Carl; 1965. "Aspects of Scientific Explanation," pp. 331-496 in his *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hirsch, Alex; 2012. "Irrational Treasure," *Gravity Falls*. Disney Channel. Television.

Hitchcock, Christopher; 2001. "The Intransitivity of Causation Revealed in Equations and Graphs," pp. 273-299 in *The Journal of Philosophy* vol. XCVII no. 6.

Hitchcock, Christopher; 2013. "What is the 'Cause' in Causal Decision Theory?" pp. 129-146 in *Erkenn* no. 78.

Hofstadter, Douglas.; 1982. "Can creativity be mechanized?" pp. 20-29 in *Scientific American* no. 247.

Humberstone, I.L.; 1992. "Direction of Fit," pp. 59-83 in *Mind* vol. 101 no. 401.

Hume, David; 1740 (2000). *A Treatise of Human Nature*. ed. David Norton and Mary Norton. Oxford: Oxford University Press.

Jeffrey, Richard; 1983. *The Logic of Decision*, 2nd edition. Chicago: University of Chicago Press.

Joyce, James; 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Joyce, James; 2012. "Regret and Instability in Causal Decision Theory," pp. 123-145 in *Synthese* vol. 187.

Kaminski, Juliane; Call, Josep; Tomasello, Michael; 2008. "Chimpanzees know what others know, but not what they believe," pp. 224-234 in *Cognition* no. 109.

Kenny, A.J.; 1966. "Practical Inference," pp. 65-75 in *Analysis* vol. 26 no. 3.

Kim, Jaegwon; 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kitcher, Philip; 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, ed. P. Kitcher and W. Salmon, 410-505. Minneapolis: University of Minnesota Press.

Kolodny, Nico; 2005. "Why Be Rational?" pp. 509-563 in *Mind* vol. 114 no. 455.

Kolodny, Nico; 2007. "State or Process Requirements?" pp. 371-385 in *Mind* vol. 116 no. 462.

Korsgaard, Christine; 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Korsgaard, Christine; 2009. *Self-Constitution*. Oxford: Oxford University Press.

Knobe, Joshua; 2003. "Intentional Action and Side Effects in Ordinary Language," pp. 190-194 in *Analysis* vol. 63 no. 3.

Kornell, Nate; Son, Lisa; Terrace, Herbert; 2007. "Transfer of Metacognitive Skills and Hint-Seeking in Monkeys," pp. 64-71 in *Psychological Science* vol. 18 no. 1.

Kriegel, Uriah; 2003. "Consciousness as Intransitive Self-Consciousness," pp. 103-132 in *Canadian Journal of Philosophy* vol. 33 no. 1.

Langton, Rae; 2004. "Intention as Faith," pp. 243-258 in *Royal Institute of Philosophy Supplement* vol. 55.

Lavin, Doug; 2013. "Must There Be Basic Action?" pp. 1-13 in *Noûs* vol. 47 no. 2

Leslie, Alan; Knobe, Joshua; Cohen, Adam; 2006. "Acting Intentionally and the Side-Effect Effect," pp. 421-427 in *Psychological Science* vol. 17 no. 5.

Lewis, David; 1979a. "Attitudes De Dicto and De Se," pp. 513-543 in *The Philosophical Review* vol. 88 no. 4.

Lewis, David; 1979b. "Counterfactual Dependence and Time's Arrow." *Noûs* vol. 13 no. 4.

Lewis, David; 1983. "Individuation by Acquaintance and by Stipulation," pp. 3-32 in *The Philosophical Review* vol. 92 no. 1.

Lewis, David; 1986. "Causal Explanation," pp. 214-240 in *Philosophical Papers* vol. 2. Oxford: Oxford University Press.

Li, Bihui. "The Evolution of Rigor, Syntax and Semantics in Applied Mathematics." Unpublished.

Mach, Ernst; 1897 (1914). *The Analysis of Sensations*. Trans. C.M. Williams, revised Sydney Waterlow. Chicago: Open Court Publishing.

McDowell, John. "How Receptive Knowledge Relates to Practical Knowledge," unpublished.

Mele, Alfred; 1992. *Springs of Action*. Oxford: Oxford University Press.

Mele, Alfred; Moser, Paul; 1994. "Intentional Action," pp. 39-68 in *Noûs* vol. 28 no. 1.

Moran, Richard; 2001. *Authority and Estrangement*. Princeton: Princeton University Press.

Moran, Richard; 2004. "Anscombe on Practical Knowledge," pp. 43-68 in *Agency and Action (Royal Institute of Philosophy Supplement no. 55)*. Cambridge: Cambridge University Press.

Nadelhoffer, Thomas; 2006. "Desire, Foresight, Intentions, and Intentional Actions," pp. 133-157 in *Journal of Cognition and Culture* vol. 6 no. 1-2.

Nozick, Robert; 1969. "Newcomb's Problem and Two Principles of Choice," pp. 114-146 in *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher. Springer.

Parfit, Derek; 1986. *Reasons and Persons*. Oxford: Oxford University Press.

Paul, L.A.; Hall, Ned; 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Paul, Sarah; 2009a. "How We Know What We're Doing," pp. 1-24 in *Philosopher's Imprint* vol. 9 no. 11.

Paul, Sarah; 2009b. "Intention, Belief, and Wishful Thinking," pp. 546-557 in *Ethics* vol. 119 no. 3.

Paul, Sarah; 2010. "Deviant Formal Causation," pp. 1-23 in *Journal of Ethics and Social Philosophy* vol. 5 no. 3.

Paul, Sarah; 2012. "How We Know What We Intend," pp. 327-346 in *Philosophical Studies* vol. 161 no. 2.

Pearl, Judea; 2000. *Causation*, 2nd edition. Cambridge: Cambridge University Press.

Perry, John; 1977. "Frege on Demonstratives," pp. 474-497 in *The Philosophical Review* vol. 86 no. 4.

Perry, John; 1979. "The Problem of the Essential Indexical," pp. 3-21 in *Noûs* vol. 13 no. 1.

Perry, John; 1993. "Perception, Action, and the Structure of Believing," pp. 121-150 in his *The Problem of the Essential Indexical*. Oxford: Oxford University Press.

Peterson, Candida; Wellman, Henry; Liu, David; 2005. "Steps in Theory of Mind Development for children with deafness or autism," pp. 502-517 in *Child Development* vol. 76 no. 2.

Price, Huw; 1986. "Against Causal Decision Theory," pp. 195-212 in *Synthese* vol. 67.

Price, Huw; 1993. "The Direction of Causation," pp. 253-267 in *PSA 1992* ed. D. Hull, M Forves, and K. Okruhlik. Vol. 2. East Lansing: Philosophy of Science Association.

Rachels, James; 1975 (2009). "Active and Passive Euthanasia," pp. 212-218 in *Exploring Ethics* ed. Steven Cahn. Oxford: Oxford University Press.

Railton, Peter; 1981. "Probability, Explanation, and Information." *Synthese* 48*:* 233-56.

Raz, Joseph; 2005. "The Myth of Instrumental Rationality," pp. 2-28 in *The Journal of Ethics and Social Philosophy* vol. 1 no. 1.

Rödl, Sebastian; 2005 (2012). *Categories of the Temporal*. Trans. Sibylle Salewski. Cambridge: Harvard University Press.

Ryle, Gilbert; 1949. *The Concept of Mind*. Chicago: University of Chicago Press.

Schaffer, Jonathan; 2010. "The Least Discerning and Most Promiscuous Truthmaker", *Philosophical Quarterly*, 69: 307–324.

Schaffer, Jonathan; 2015. "Grounding in the Image of Causation," *Philosophical Studies*. DOI: 10.1007/s11098-014-0438-1.

Schelling, Thomas; 1981. *The Strategy of Conflict*. Cambridge: Harvard University Press.

Schroeder, Mark; 2004. "The Scope of Instrumental Reason," pp. 337-364 in *Philosophical Perspectives* vol. 18.

Searle, John; 1983. *Intentionality*. Cambridge: Cambridge University Press.

Setiya, Kieran; 2007a. *Reasons without Rationalism*. Princeton: Princeton University Press.

Setiya, Kieran; 2007b. "Cognitivism about Instrumental Reason," pp. 649-673 in *Ethics* vol. 117 no. 4.

Setiya, Kieran; 2008. "Practical Knowledge," pp. 388-409 in *Ethics* vol. 118 no. 3.

Setiya, Kieran; 2009. "Practical Knowledge Revisited," pp. 128-137 in *Ethics* vol. 120 no. 1.

Setiya, Kieran; 2012. "Knowing How," pp. 285-307 in *Proceedings of the Aristotelian Society* vol. CXII part 3.

Sextus Empiricus; 2000. *Outlines of Skepticism*. ed. Julia Annas, trans. Jonathan Barnes. Cambridge: Cambridge University Press.

Shoemaker, Sydney; 2003. "Moran on Self-Knowledge," pp. 391-401 in *European Journal of Philosophy* vol. 11 no. 3.

Sider, Theodore; 2012, *Writing the Book of the World*, Oxford: Oxford University Press.

Spirtes, Peter; Glymour, Clark; and Scheines, Richard; 2001. *Causation, Prediction, and Search*, 2nd edition. Cambridge: The MIT Press.

Thalberg, Irving; 1984. "Do Our Intentions Cause Our Intentional Actions?" pp. 249-260 in *American Philosophical Quarterly* no. 21.

Thompson, Michael; 2008. *Life and Action*. Harvard: Harvard University Press.

Thomson, Judith Jarvis; 1971. "A Defense of Abortion," pp. 47-66 in *Philosophy and Public Affairs* vol. 1 no. 1.

Velleman, J. David; 1989 (2000). "Epistemic Freedom," pp. 32-55 in his *The Possibility of Practical Reason*. Oxford: Oxford University Press.

Velleman, J. David; 1996 (2000). "The Possibility of Practical Reason," pp. 170-199 in his *The Possibility of Practical Reason*. Oxford: Oxford University Press.

Velleman, J. David; 2007. *Practical Reflection*. CSLI Publications.

Wallace, R. Jay; 2001. "Normativity, Commitment, and Instrumental Reason," pp. 1-26 in *Philosopher's Imprint* vol. 1 no. 3.

Wedgwood, Ralph; 2006. "The Normative Force of Reasoning," pp. 660-686 in *Noûs* vol. 40 no. 4.

Weirich, Paul; 2012. "Causal Decision Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/decision-causal/>.

Wellman, Henry; Liu, David; 2004. "Scaling of Theory-of-Mind Tasks," pp. 523-541 in *Child Development* vol. 75 no. 2.

Wellman, Henry; Fang, Fuxi; Liu, David; Zhu, Liqi; Liu, Guoxiong; 2006. "Scaling of Theory-of-Mind Understandings in Chinese Children," pp. 1075-1081 in *Psychological Science* vo. 17 no. 12.

Wellman, Henry; Fuxi, Fang; Peterson, Candida; 2011. "Sequential Progressions in a Theory-of-Mind Scale," pp. 780-792 in *Child Development* vol. 82 no. 3.

Williams, Bernard; 1970 (1973). "Deciding to Believe," pp. 136-151 in his *Problems of the Self*. Cambridge: Cambridge University Press.

Wilson, George; 1989. *The Intentionality of Human Action*. Stanford: Stanford University Press.

Woodward, James; 2003. *Making Things Happen*. Oxford: Oxford University Press.

Woodward, James; 2014. "Interventionism and Causal Exclusion," *Philosophy and Phenomenological Research*. DOI: 10.1111/phpr.12095

Wooldridge, Dean; 1963. *The Machinery of the Brain*. New York: McGraw Hill.

Wilson, Mark; 2006. *Wandering Significance*. Oxford: Clarendon Press.