FROM KNOWLEDGE OF OUR EXISTENCE TO NORMATIVE KNOWLEDGE

Michael Smith

1. Background

One of the main questions on which analytic philosophers disagree is whether there is a fact-value gap, or, equivalently, whether we can derive 'ought' claims from 'is' claims. Hume famously thought that there is such a gap. What we do do is one thing; what we ought to do is quite another; and no amount of knowledge and understanding of the former suffices for knowledge and understanding of the latter. In what follows I explain how and why we should resist Hume's conclusion. More specifically, I will argue that certain claims about what we ought to do are derivable from some fundamental a priori truths about our nature, truths that are manifestly factual.

The strategy of argument to be pursued in what follows has an ancient pedigree, as it is the same as the strategy Aristotle pursued in his famous Function Argument (1984). The strategy requires that we first identify what Judith Jarvis Thomson calls a "goodness-fixing kind" of which we are members (Thomson 2010). A goodness-fixing kind is any kind the understanding of which allows us to order members of that kind from best to worst. The best examples of goodness-fixing kinds are functional kinds. The constitutive feature of a heart, for example, is that it is an organ whose function is to pump an adequate supply of blood around the body. Because this is the constitutive feature of hearts, it follows that we can order hearts from those that succeed in pumping an adequate supply of blood around the body (these are the best) to those that do so fairly well (these are slightly worse than the best), to those that barely do so at all (these are significantly worse than the best), to those that utterly fail to do so (these are the worst).

Note that this ordering from best to worst can be understood in purely descriptive terms. This is the key to the argument that follows, as equipped with this ordering, we can similarly understand in purely descriptive terms a whole range of other normative terms, both evaluative and deontic. Imagine that you have aching limbs and cold extremities, so you go to the doctor. He examines you and then wonders aloud, 'Why isn't your heart functioning properly?' or, 'Why isn't your heart doing what it is supposed to do?' or 'Why is your heart defective?' We can explain the doctor's use of 'properly', 'supposed to', and 'defective' in terms of the ordering. The constitutive feature of a heart is that its function is to pump an adequate supply of blood around the body (this is an initial *is*- or *factual*-claim), which generates the ordering (an initial set of *evaluative* claims), which in turn entails that a *properly* functioning heart is one that does pump an adequate supply of blood around the body (an initial *deontic* claim), that hearts are *supposed to* pump adequate supplies of blood around the body (another *deontic* claim), that a *defective* heart is one that fails to do so (another *evaluative* claim), and so on. There is therefore no quite general *is-ought* or *fact-value* gap.

As I said above, the strategy of argument in what follows requires us to find some goodness-fixing kind of which we are members, but it doesn't just require that. It requires in addition that that knowledge that we are members of that goodness-fixing kind is suitable for giving us knowledge of the kinds of normative claims in which we were originally interested. It is this second requirement that has proven to be the stumbling block of earlier attempts to pursue the strategy. To give just one example, when Philippa Foot pursues the strategy in her *Natural Goodness* (2001), she focuses on our being members of the species-kind *human being*. The ought-claims she derives from our being members of this biological kind thus have the same epistemic status as our knowledge that we are human beings: that is, they are all a posteriori. If, as it seems to me, the ought-claims in which we are interested are knowable a priori—the knowledge of the kind of ought-claims in which we are interested is *armchair knowledge*—then Foot's pursuit of the strategy goes wrong at precisely this point. The strategy of argument pursued in what follows aims to avoid going wrong in this way.

Email: msmith@princeton.edu

2. Knowledge from the armchair that we are thinkers in space-time

Let's start at the very beginning. Are there any ways of thinking about ourselves and our place in nature that are rationally obligatory, and so could reasonably be expected to be shared by anyone capable of thinking at all, provided they thought things through? Descartes famously thought that there were, and so we will begin with what he should have concluded about himself and the world he inhabits when he tried to figure out from the armchair whether there is anything that he knew (Descartes 1637-1642, Williams 1978).

There is, I think, an important lesson to be learned by starting with Descartes' famous thought experiment, but in order to learn that lesson we need to give up on much that he either took for granted or believed because of what he took for granted. Most importantly, we must give up on the idea that *indubitability* is the standard that a claim needs to meet in order to count as known a priori. If, as seems plain, we know a great many claims a priori on the basis of inference to the best explanation, the fact that we could bring ourselves to doubt such claims must be neither here nor there. Doubts can be unreasonable, and doubting something that is part of the best explanation of something that requires an explanation is a paradigmatic case of being unreasonable. As we will see, this opens terrain that Descartes didn't get to explore.

So let's imagine ourselves in Descartes' position, but let's do so while thinking more like a modern metaphysician. We want to figure out whether anything can be known to be true, and, at least initially, we will try not to rely on any empirical information beyond what's available to us simply by engaging in the activity itself. We will abstract away from the fact that we are sitting in an armchair, and even the fact that we are human, and so on, and we will only allow ourselves to dwell on the question. But from the fact that we are doing that, and thereby entertaining various thoughts, together with the premise that thoughts require thinkers, we can conclude that there is something that can be known, namely, that we, the thinker of these very thoughts, exist. That we exist is clearly a contingent fact, but we appear to know it a priori, as we know it to be true simply by understanding what the thought is that we are entertaining. There is a great deal packed in here, and it is worth going through it slowly to make sure we identify the different elements.

For one thing, the process of thinking things through takes time, so the conclusion that we exist concerns not just a presently existing entity, but a persisting entity. As a thinker we existed in the immediate past when we set ourselves the task, and we persisted as we continued thinking things through, and our persistence explains why we exist in the present. It therefore seems that we are in a position to know not just one thing, but two things: that we exist, and that so too does a temporal order. What the nature of that temporal order is—whether we can make sense of time's arrow, for example (Price 1996)—is entirely up for grabs. The point is just that we are committed to thinking that a temporal order of some sort exists.

With those thoughts in mind, two further possibilities need to be distinguished. When we think of ourselves as thinkers, are we supposing, here in the present, that if we existed in the past, then we were thinking in the past? Or are we supposing that we could have existed in the past without thinking in the past? These questions force us to ask whether in thinking of ourselves as thinkers, we are thinking of ourselves as thinking essentially; or whether we are instead thinking of ourselves as beings with the capacity to think, a capacity that we may exercise at certain times but not others; or whether we are thinking of ourselves as beings with some other nature entirely, beings who could have existed and continue to exist even when we have lost our capacity to think.

The answer to the first of these questions, at least, once again seems to be clear. There are discontinuities in our memories of our thinking, and the best explanation of those discontinuities requires us to suppose that if we do think, then we do so because we exercise our capacity to think. Since we have this capacity, it follows that we might always exercise our capacity to think. But even if we did always exercise that

capacity, we can still conceive of the possibility that we do not, and this is the possibility that seems to fit best with our memories of discontinuities in our thinking. We had the capacity to think in the past, but we sometimes failed to exercise that capacity, and this explains the discontinuities in our memories. Moreover we can expect that the same will be true about our thinking in the future.

It therefore seems that we are in a position to know not just two things, but three things: that we exist, that so too does a temporal order, and that what we are within that temporal order is the ground of our capacity to think, whatever that may be, where this is a capacity that we may or may not exercise. Note that at this point we have well and truly parted company with Descartes. Whereas he thought that the ground of the capacity to think itself had to be a mental substance, we haven't seen any reason to suppose that this to be so. Moreover whether *being with the capacity to think* is a phase sortal that we fall under, like *infant*, and *adult*, or our substance sortal, which is to say the kind of thing whose coming into and going out of existence is our coming into and going out of existence, is still up for grabs. More on this presently.

We have just seen that we can conceive of ourselves as existing, but without thinking. Can we also conceive of there being thinking, but where that thinking is not ours? In other words, can we conceive of ourselves not being alone? The answer once again seems to be that we can, though this is more contentious. There are other thinkers just in case there are grounds of the capacity to think other than the grounds of our own capacity to think. For there to be such grounds, there would have to be the means to distinguish the ground of our capacity to think from the ground of others' capacities to think, and there would have to be the means by which to think of them as co-existing, not just in the sense of both being possible, but in the sense of both being, for all we know, actual. This suggests that there would have to be spatial dimensions in which to locate the different ground of our own capacity to think and the grounds of others' capacities to think, if such grounds exist. To co-exist with other thinkers would then amount to no more or less than bearing spatio-temporal relations to them.

But now consider what it would be for us to exist but be alone. One way for us to exist but be alone would be for there to be a possible world which is otherwise just like the possible world in which we are thinking along with other thinkers, but in which there are no other thinkers, just us. This is a possible world in which is there is still a spatio-temporal order, albeit one in which we are located as the only thinker. The other way is for there to be a possible world in which there are no other thinkers, and no spatial order either, just a temporal order. A principle of parsimony tells against there being the latter possible world. There is only one way of being alone. Possible worlds are one and all spatio-temporally ordered.

Now consider a possible world, which is to say a spatio-temporal order, in which we are located as the only thinker. At first glance it seems that there are two further ways in which we could be located in that spatio-temporal order as the only thinker. One is for there to be no possible experience that would support the conclusion that, at the very moment at which we are thinking that p, there is also thinking that not p. But since there plainly are such experiences—we have such experiences, but interpret them as providing evidence that we are not alone—we can ignore this. The other is for there to be possible such experiences, but for them to suggest not that we are thinking that p in one spatial location, and that someone else is thinking that not p in another, but rather that we have spatial parts, and that we can simultaneously be thinking that p in one spatial part and not p in another.

I said earlier that it is contentious to suppose that we are not alone. It is contentious because, even supposing that we do have experiences that support the conclusion that at the very moment at which we are thinking that p, there is also thinking that not p, there are two interpretations of that evidence. One is that we are not alone, and that another thinker is simultaneously located somewhere else. The other is that we have spatial parts, and that we can simultaneously have contradictory thoughts in our different parts. The upshot is that we are in a position to know not just three things,

but four things: that we exist, that so too does a temporal order, that so too does a spatial order, and that we are located within that spatio-temporal order as the ground of the capacity to think, whatever that may be, and that, with the caveat just mentioned, other agents may be so located too (for more on this see Johnston 2010).

As with the temporal order, the nature of that spatial order is still up for grabs. Should we think of the spatial order in absolute or relational terms? Nothing we have said so far tells in favor of one or another of these. The point is just that we are committed to thinking that a spatial order of some sort exists.

3. Knowledge from the armchair that we are agents

So far I have suggested that we are the ground of our capacity to think. But remember how this whole thought experiment began. We wanted to figure out whether there is anything that can be known, and we have concluded that we know quite a lot. We were curious. We desired to figure out whether we know anything, and we satisfied that desire by thinking and thereby acquiring knowledge. This suggests that we are also the ground of the capacity to know things about the world in which we live, and to satisfy our desires in that world. In other words, we are *agents*, where an agent is anything with the two capacities just mentioned: that is, anything with the capacity to know the world in which it lives and satisfy its desires in that world.

The class of agents thus includes humans like us who are capable of rehearsing for ourselves Descartes' thought experiment, and so figuring out what they are, but if there are ways of satisfying desires and gaining knowledge without having any thoughts, then the class of agents also includes less sophisticated beings that may be incapable of thought. To return to a question that we left hanging a few moments ago, this raises the possibility that the capacity to think is not essential to us. Perhaps *being with the capacity to think* is a phase sortal after all, and a better candidate for our substance sortal is *agent*. We won't take a stand on this question here, but we will return to it later.

To be clear, what actions did we perform when we imagined ourselves in Descartes position? The answer is that we performed various mental actions. We desired to figure out whether we knew anything, we believed that the way to do that was by (say) focusing our attention on the question and thinking it through, and that desire and belief led us to desire to focus our attention on the question, which desire in turn led us to focus our attention, and having focused our attention on the question, our thought processes took off. There are various candidates for the mental actions we performed here, as we did several things: one thing we did was to form the desire to focus our attention, another was to focus our attention, yet another was to think various thoughts, and yet another was to figure out that we know something. We will talk more about forming desires and thinking thoughts later on, but for the moment let's fix on focusing our attention and figuring out that we know something.

What makes focusing our attention and figuring out that we know something especially good candidates for mental actions we performed is that, in the circumstances, they both seem to be things that we did intentionally. We didn't focus our attention as the result of external forces, as we might do if someone shouted our name, or there was a loud bang. We focused our attention as a way to get something we wanted, that is, as a way to figure out whether we know something. Moreover, focusing our attention seems to be a *basic* mental action, that is, a mental action that we know how to perform, but where our knowledge of how to perform that action isn't explained by our knowledge of how to perform some other action (compare Danto 1965). Contrast the mental action of focusing our attention with the mental action that we performed by doing that, namely, figuring out whether we know something. The latter is something we know how to do, but only because we know how to do other things that have our having figured out whether we know something as their upshot, things like focusing our attention and then thinking the matter through, a process that results in our knowing something.

I have labored the point about mental actions for a number of reasons. One is to bring out the point just made. There are differences in our knowledge of how to perform mental actions: some are basic and others are non-basic. Another is to bring out how saturated the world as understood from the armchair is with causes and effects, where at least some of these, the mental actions that we know how to perform without performing some other mental action, are under our intentional control. Our desire to figure out whether we know anything and our belief that we could do that by focusing our attention *led us* to desire to focus our attention on the question, that in turn *led us* to focus our attention, and *the upshot* of our focusing our attention and the thinking that ensued from the exercise of our capacity to think was the knowledge that we exist.

It seems that we are therefore in a position to know not just four things from the armchair, but five things: that we exist, that so too does a temporal order, that so too does a spatial order, that we are located within that spatio-temporal order as the ground of the capacity to think, whatever that may be, and that others may be so located too; and that elements within that spatio-temporal order bear causal relations to each other, some of which are under our control, and hence that, even more fundamentally, we are the ground of the capacity to gain knowledge of the world and realize our desires in it. As with the nature of the spatio-temporal order, the nature of the causal relation is still up for grabs. The point is just that we seem to be committed to the existence of causal relations, conceived of in some way or other.

With this much knowledge of ourselves as agents from the armchair, note that we are primed to understand what our agency consists in when we leave the armchair. This is because we know from the armchair that if our beliefs were causally connected to elements in the spatio-temporal order in such a way as to give us knowledge of those elements, and if we had desires about how the elements in the spatio-temporal order are to be, and hence knowledge of the changes that would need to be made, if changes need to be made, in order for those elements to be that way, and if these desires and beliefs were connected causally to some element in the spatio-temporal order that was under our control such as a body—that is to say, an element we knew how to change, where our knowledge of how to change that element isn't explained by our knowledge of how to change any other element—then we would be in a position to intentionally make changes to that element—that is to say, move our bodies—that would in turn cause the changes in the elements that we desire. In other words, we would be in a position to have knowledge of the elements of the spatio-temporal order and to intentionally make changes to them so as to realize our desires.

When we leave the armchair, what we discover is of course that these conditions are met. We have sense organs that are causally connected to, and so provide us with knowledge of everyday objects in the spatio-temporal order—things like other people, food, shelter, and clothing—and we have a body that is causally connected to those objects, and we have knowledge of how to move that body, where that knowledge isn't explained by our knowledge of how to do anything else. We have desires about how the objects in the spatio-temporal order are to be—that is, we have desires for things like the company of other people, food, shelter, and clothing—and these desires combine without our knowledge of how to move our body and our knowledge of the changes that the movements of our body would bring about to the objects in the world to produce those movements, which in turn bring about those changes. In other words, we know where to find the people whose company we seek, and the food, shelter, and clothing we need, and we act intentionally so as to secure these things for ourselves.

The story of non-mental action just told should sound familiar, as it is the so-called 'standard story of action' that we inherited from Aristotle and Hume, and that was developed and refined by Donald Davidson in the 1960s and 70s (from here-on I will omit the qualifier 'non-mental'). Under Davidson's influence, we have come to think of actions as having certain constitutive features. An action is just a bodily movement

that is caused in the right way by a desire for an end—that is, an intrinsic desire—and a means-end belief, a view of non-mental action that we can represent as follows:

means-ends belief

→ bodily movement = action

desire for an end

In this picture, the '→' represents the fact that the right kind of causal relation holds between the means-end belief and the desire for an end, on the one hand, and the bodily movement, on the other (Davidson 1963, 1971a, 1971b, 1973, 1976). But though this Davidsonian way of thinking about action has been extremely influential, it is in fact far too simple. It needs to be modified in ways suggested by two other defenders of the standard story, Davidson's teacher Carl Hempel (1961) and his student Michael Bratman (1987).

Let's begin with Bratman's modification. What happens, according to this way of understanding the standard story, if you have just one desire for an end, but two means-end beliefs and nothing to choose between them? For example, imagine desperately wanting some food for breakfast, and being confronted by a supermarket shelf with many identical packets of muesli on it. There is nothing you will get from taking the muesli packet immediately in front of you but slightly to the left that you won't get from taking the one immediately in front of you but slightly to the right, and vice-versa, but there is definitely something that you will get from taking one or the other that you won't get from taking any of the other options available to you, namely, breakfast. Taking one or the other is the most efficient way of getting what you desperately want, but there is nothing that tells in favor of choosing one of these rather than the other. What explains your choice, in these circumstances? Davidson's version of the standard story can't answer that question.

What's needed in these circumstances, as Bratman points out, is a way of forming some further mental attitude that fixes you on one of your fully determinate options and keeps you fixed on that option unless there are significant changes in the background desires for ends and means-end beliefs in the context of which that further mental attitude was formed. Bratman thinks of this further mental attitude as an intention, but we might equally well think of it, as Richard Holton does in his development of Bratman's view, as a willing—that is, as the upshot of the exercise of a power that we possess to will one of our options in such circumstances—so the modified picture looks like this:

means-ends belief

(→ willing) → bodily movement = action

desire for an end

A willing can be understood as a state that takes as inputs desires for ends and means-end beliefs that leave an agent indifferent between a range of options, and that has as an output the agent's acting as if he desires most to pursue one of those options. The brackets thus represent the fact that the intention or willing may only be necessary in cases where the combination of the desire for an end and a means-end belief underdetermine the choice.

The other modification is more substantial and requires more argument. Two years before Davidson published his version of the standard story of action, his teacher Hempel had published his, but Hempel's version, though similar to Davidson's, is different in a crucial respect. Hempel thought that an additional psychological element is required among the causes of a bodily movement for that movement to count as an action. To see why Hempel thought this, imagine an agent, John, who has an intrinsic desire to get stronger and the belief that something he can just do, namely lift weights, would make him stronger. According to Hempel, what's needed for John to lift weights isn't just the intrinsic desire and means-end belief, but also the capacity

to be instrumentally rational in the circumstances and the exercise of that capacity. This is needed because if John is instrumentally irrational, then notwithstanding the fact that he has an intrinsic desire to get stronger and a belief that he can get stronger by lifting weights, he will not form the instrumental desire to lift weights, and without that instrumental desire, he will not lift weights.

We can represent the standard story with both Bratman's and Hempel's modifications as follows:

means-ends belief

+ (→ willing) → bodily movement **=** action

desire for an end

In this picture, the '**+**' represents the agent's possession and exercise of the capacity to be instrumentally rational as an additional causal element. Its presence guarantees that the agent puts his desire for an end and means-end belief together so as to make it the case that he has a desire for the means. Indeed, on this picture instrumental desires—that is to say, desires for means—are best thought of as nothing over and above intrinsic desires and means-end beliefs that have been put together by the agent's exercise of his capacity to be instrumentally rational.

There is, I think, an important truth lurking here, but in order to see what that truth is it will be helpful to imagine how Davidson might reply to Hempel. The issue, Davidson might say, is whether the agent's being instrumentally rational in the circumstances is already guaranteed in his sparser story by the requirement that, when an agent acts, the intrinsic desire and means-end belief *cause* the bodily movement *in the right way*. In other words, he might reply, the **+** in Hempel's more complicated story is guaranteed by the → in his? But is this reply successful? Let's focus on Davidson's requirement that, for an agent to act, his bodily movements must be caused *in the right way* by his intrinsic desires and mean-end beliefs.

Imagine an actor playing a role that calls for her to shake as if extremely nervous. We can readily suppose that, despite the fact that she wants to play her role and believes that she can do so by shaking, once she gets on stage her desire and belief so unnerve her that she is overcome and rendered totally incapable of action. Instead of playing her role as required, she just stands there, shaking nervously. She doesn't act, but instead something happens to her. It was because of examples like this that Davidson concluded that it is insufficient for an agent's bodily movements to be actions that she has relevant desires and beliefs that cause those movements. He thought, correctly, that we need to rule out the possibility of her desires and beliefs causing her to shake in the wrong way, for example, via causing her to become nervous.

Davidson himself was pessimistic about the possibility of our ruling that possibility out in anything other than an uninformative way—that is, by stipulating that the relevant desires and beliefs must cause the bodily movements *in the right way*—but many think that it is plain what is needed (Peacocke 1979). The crucial feature in all cases of causation in the wrong way, they say, is that the match between what the agent does and the content of her desires and beliefs is entirely fluky. In the case just described, for example, it is entirely fluky that the actor wanted to make just the movements that her nerves subsequently caused. In order to state a sufficient condition for an agent's bodily movements being actions, we must therefore ensure that her movements are especially sensitive to the content of her desires and beliefs, as opposed to being sensitive to the operation of irrelevant factors like nerves.

The movement of an agent's body is an action, the suggestion goes, only if, in addition to the conditions Davidson mentions, over a range of desires and beliefs that the agent might have had that differ ever so slightly in their content, she would still have performed an appropriate bodily movement. Had she desired to act nervously and believed that she could do so by making her teeth chatter, then she would have made her teeth chatter. Had she desired to act nervously and believed that she could

do so by walking around wringing her hands, then she would have walked around wringing her hands. And so on. This further condition of *differential sensitivity* of the agent's bodily movement to the specific contents of her intrinsic desires and means-end beliefs is clearly violated in cases of internal wayward causal chains because, to return to the example discussed above, even if the actor had had such ever-so-slightly different desires and beliefs, her nerves would still have caused her to shake when she went on stage.

Whether or not this further requirement of differential sensitivity of an agent's bodily movements to ever-so-slight variations in the contents of her desires and beliefs turns the necessary condition for a bodily movement's being an action into a necessary and sufficient condition is a moot point (see Sehon 2005). But for present purposes, that's not what's important. What's important is simply that some such differential sensitivity requirement on the relationship between an agent's bodily movements and her desires and beliefs is required. But now consider the differential sensitivity requirement itself. What does it amount to? It amounts to nothing less than the requirement that the agent has and exercises the capacity to be instrumentally rational *in a very local domain*. For a desire and belief to cause a bodily movement in the right way for that bodily movement to count as an action, is, inter alia, for the agent to have and exercise his capacity to be instrumentally rational in those circumstances.

Consider again the example just discussed. It isn't enough that the agent has the instrumental desire to shake. She must also be such that she would have had the instrumental desire to wring her hands if she had believed that wringing her hands was a way of acting nervous; that she would have had the instrumental desire to make her teeth chatter if she had believed that making her teeth chatter was a way of acting nervous; and so on for ever so slight variations in the contents of the agents beliefs and intrinsic desires. The requirement that intrinsic desires and means-end beliefs cause actions in the right way thus does indeed seem to entail that the agent has and exercises the capacity to be instrumentally rational, at least in a very local domain. So far, then, the main thrust of Davidson's imagined reply to Hempel appears to be on the mark. But a little more thought reveals that the reply misses the mark. Hempel was right and Davidson was wrong.

The capacity to be instrumentally rational whose exercise plays an explanatory role in the production of action need not be the exercise of the very localized capacity to be instrumentally rational that we have seen is necessary for action. In order to see that this is so we need to consider the various ways in which an agent's being more fully instrumentally rational in the circumstances in which he acts may and may not manifest itself, and how this differs from the manifestation conditions of the very localized capacity just described. Let's therefore begin by imagining a slightly more complicated, but still quite simple, example. Suppose that John has an intrinsic desire to get stronger and that he believes there are two ways in which he could bring this about. He believes that his getting stronger would result from lifting weights or from swimming, but he does not believe that he could lift weights and swim at the same time. If John were fully instrumentally rational, what would he desire to do in this case?

The answer is that if John were fully instrumentally rational then he would put his intrinsic desire to get stronger together with each of these beliefs. This is because his intrinsic desire is already targeted, so to speak, on each of these ways the world could be. He desires the realization of the possibility that he is strong, and he believes that this possibility partitions into two sub-possibilities: the possibility that he lifts weights and the possibility that he swims. Putting at least one of his means-end beliefs together with his intrinsic desire would allow him to be instrumentally rational to a certain degree—that would amount to a very local exercise of his capacity to be instrumentally rational—but he would be more instrumentally rational if he were to put his intrinsic desire together with both his means-end beliefs. He would be more instrumentally rational because doing so prepares him for action in a more modally robust sense: he would be such that, had he believed himself unable to (say) lift

weights, he would still have desired to swim, and vice versa. If, as seems plausible, being fully instrumentally rational is a matter of maximal preparedness to act in this modally robust sense, then being fully instrumentally rational would seem to require him to have both an instrumental desire to lift weights and an instrumental desire to swim.

Moreover, sticking with this case, John's being fully instrumentally rational would seem to have implications for the strengths of John's instrumental desires. If he is equally confident about the two causal claims just made—equally confident that lifting weights will cause him to get stronger and that swimming will cause him to get stronger—then, if he were fully instrumentally rational, he would be indifferent between the two options. His instrumental desires would be equally strong, and a willing would be required to choose between his options. But if he is more confident of one than the other, then in order to satisfy all of the demands of instrumental rationality, his instrumental desire for the one about which he is more confident would have to be stronger. The effect of decreased confidence should be to dilute the instrumental desire for that option. This too manifests itself modally. If John is fully instrumentally rational then he is actually such that he instrumentally desires more that about which he is more confident, but had he believed that to be impossible, he would have instrumentally desired that about which he is less confident. So even though agents might be instrumentally rational to the extent that their intrinsic desires are suitably related to two means-end beliefs they have, they might still fail to meet instrumental rationality's further demand on the strengths of their two instrumental desires.

Instrumental rationality would seem to make other more global demands on agent's instrumental desires as well. Suppose this time that John has two desires, an intrinsic desire to get stronger and an intrinsic desire for knowledge, and that he believes all of the following: that lifting weights causes strength, that reading causes knowledge, and that he cannot lift weights and read at the same time—the mental energy required to lift weights is inconsistent with his simultaneously reading for understanding. Finally, just to keep things simple, suppose he is equally confident about each of these and that he has no further desires or beliefs. If John were fully instrumentally rational, then the above considerations would apply equally to the two intrinsic desires. Instrumental rationality requires that his two intrinsic desires be suitably related to each of his means-end beliefs. If he were instrumentally rational then he would have both an instrumental desire to lift weights and an instrumental desire to read.

Moreover it also seems that, though he might be instrumentally rational in this local sense, he might fail to meet a further demand that instrumental rationality makes on the strengths of these instrumental desires. If his intrinsic desires for strength and knowledge are equally strong then, if he were instrumentally rational, he would be indifferent between the two options: his instrumental desires to lift weights and to read would be equally strong. A willing would be required to choose between his options. But if one of his intrinsic desires is stronger than the other then in order to satisfy the more global demands of instrumental rationality, his instrumental desire for the one which leads to the outcome that he desires more strongly would have to be stronger. The effect of having one desire stronger than another in the face of equal confidence about the ways in which those desires can be satisfied should be to intensify the strength of the instrumental desire for the means to that which one desires more.

There are also cases that contain elements of both those discussed thus far. Suppose John has a stronger intrinsic desire to get stronger and a weaker intrinsic desire for knowledge, and that he believes that lifting weights causes strength, that reading causes knowledge, and that he cannot lift weights and read at the same time, but that he is more confident of the connection between reading and knowledge than he is about the connection between lifting weights and strength. What does instrumental rationality require in that case? Once again, it seems that if John were fully instrumentally rational then he would have instrumental desires both to lift weights

and to read, where the strengths of these instrumental desires would depend on the strengths of his two intrinsic desires and the levels of confidence associated with his two means-end beliefs. Indeed, if his confidence is greater enough, then instrumental rationality may even require that the instrumental desire to read is stronger than the instrumental desire to lift weights, notwithstanding the fact that the intrinsic desire for knowledge that partially constitutes it is weaker than the intrinsic desire for strength which partially constitutes the instrumental desire to lift weights.

Let's now return to the reply to Hempel that we imagined on Davidson's behalf, the reply that there is nothing for an agent's being locally instrumentally rational in the circumstances to amount to beyond the fact that his desires and means-end beliefs cause a bodily movement in the right way. We can now see that, even when an agent's desires and means-end beliefs do issue in action, and hence the agent is instrumentally rational to some extent—the agent has and exercises his capacity for instrumental rationality in the very localized domain entailed by causation in the right way—there are at least two quite distinct ways the agent might be counterfactually. This is because being instrumentally rational comes in degrees, and these two possibilities turn on the degree to which the agent is instrumentally rational in the circumstances.

Sticking with our very simple example, suppose that John has an intrinsic desire to get stronger and that he believes both that he could get stronger by lifting weights and by swimming, but that he is more confident of the former than the latter and hence, because he is instrumentally rational to a certain extent and has no other desires and means-end beliefs, he has a stronger instrumental desire to lift weights and so lifts weights. From this description of the case we cannot tell how strong John's instrumental desire to lift weights is. We know that it is stronger than his instrumental desire to swim, but that doesn't entail it is as strong as it should be, if he were fully instrumentally rational, for that requires that the strength of his instrumental desire to lift weights reflects the strength of both his intrinsic desire to get stronger and his confidence that lifting weights will lead to his getting stronger. So far all we know is that it reflects his degrees of confidence. What does this further difference consist in?

The answer is that it consists in facts about what (say) John would have done if he had also had a weaker intrinsic desire for knowledge, but had had the same level of confidence that reading a book would provide him with knowledge as that lifting weights would make him strong. One answer is that, since John's instrumental desire to lift weights would have been stronger than his instrumental desire to read a book, he would still have lifted weights. Another answer is that, since his instrumental desire to lift weights would have been weaker than his instrumental desire to read a book, he would have read a book. If the answer is the first then, in the actual circumstances, it follows that John is more globally instrumentally rational than he is if the answer is the second. This is because, if the answer is the first, the strength of his instrumental desire to lift weights reflects not just his confidence levels about the effect of lifting weights and swimming on his strength, but also the strength of his intrinsic desire to get stronger. If the answer is the second, this is not so.

We are now in a position to see why it is quite wrong to think that, since being instrumentally rational in a very local domain is entailed by an agent's desires and means-end beliefs causing his bodily movement in the right way, it follows that his being instrumentally rational cannot be a part of the explanation of his action. Different agents possess the capacity to be instrumentally rational to very different degrees, and the degree to which they possess this capacity, and whether or not they exercise their capacity to whatever extent they have it, fixes not just what actually happens when they act—fixes not just that they do exercise their capacity to be instrumentally rational in the very local domain—but also what they would do in various counterfactual circumstances, circumstances in which they have very different intrinsic desires, or in which their beliefs about their options are very different.

It is thus an agent's possession and exercise of his capacity to be instrumentally rational *to the specific extent that he has and exercises that capacity* that figures in the

explanation of his actions. To be sure, some agents may be so minimally instrumentally rational that, when they act, they thereby exercise all of the capacity to be instrumentally rational that they have. This is, if you like, the limit case of an agent. But not all agents are the limit case. Some are far more instrumentally rational than that, and, when they act, they exercise their far more extensive capacity to be instrumentally rational. This more extensive capacity is what's involved in the explanation of their actions. This is evident from the very different counterfactuals that are true of them.

What is thus true, of course is that the *minimum required* for a bodily movement to be an action is that the agent possesses and exercises the very local capacity for instrumental rationality required for his intrinsic desires and means-end beliefs to cause his bodily movement in the right way. But it would be a fallacy to move from this to the conclusion that it is an agent's possession and exercise the minimal capacity that figures in the explanation of his actions. It would be a fallacy on a par with supposing that, just because all that is strictly necessary for an agent to intentionally flip a switch (say) is that he has a very specific desire concerning the outcome of his flipping the switch, so the only desires that are ever part of the explanation of any agent's flippings of switches are desires with very specific contents.

Let me sum up. The story of non-mental action that emerges from the armchair is the standard story of action that we inherited from Aristotle and Hume, a story that was developed and refined by Davidson in the 1960s and 70s, but that requires modifications suggested by Hempel prior to Davidson and Bratman since:

means-ends belief

+ (→ willing) → bodily movement **=** action

desire for an end

According to this story, non-mental actions have certain constitutive features. They are bodily movements that stand in the right kind of causal relation to three psychological states of an agent—his intrinsic desires, his means-end beliefs, and his exercise of his capacity to be instrumentally rational—and perhaps also a willing.

Moreover, if we substitute 'mental doing' for 'bodily movement', we can see that what's true of non-mental actions is true of mental actions as well. They are just mental doings that are caused in the right way by intrinsic desires, means-end beliefs, and exercises of the capacity to be instrumentally rational, with perhaps a willing as a further causal intermediary. It seems that we are therefore in a position to know not just five things from the armchair, but six things: that we exist; that so too does a temporal order; that so too does a spatial order; that we are located within that spatio-temporal order as the ground of the capacity to think, whatever that may be, and that others may be so located too; that elements within that spatio-temporal order bear causal relations to each other, some of which are under our control, and hence that, even more fundamentally, we are the ground of the capacity to gain knowledge of the world and realize our desires in it; and that this presupposes that we are also the ground of the capacity to will and to be instrumentally rational to some extent.

The upshot is that there is a seamless transition from knowledge that we think and therefore exist, to knowledge of the kind of thing that we are. We are *agents* of the kind described in a spatio-temporal order.

4. Knowledge from the armchair of normative truths

Let's return to the main line of argument. The kind *agent* is another example of a goodness-fixing kind. The constitutive feature of agents, at least as we have understood them so far, is that they are the grounds of the capacities to realize their intrinsic desires, to have knowledge of the world in which they live, to be instrumentally rational to some extent, and to will. If we can understand each of these ideas in purely descriptive terms, as it seems we can, given the preceding discussion, then it follows that we can similarly order *agents* from best to worst in purely

descriptive terms, and these orderings will in turn entail various evaluative and ought claims.

Ideal agents are those at the top of this ranking (this is one evaluative claim), those at the bottom of the ranking who are barely agents at all are very *poor* agents (this is another evaluative claim), and those in between are *good to a certain extent* or *bad to a certain extent* (these are further evaluative claims). We can then use these evaluative claims to define claims about what agents *ought to do*, and what they *have reason to do* (these are deontic claims). To give just one example, we can understand what agents have reasons to do in terms of their ideal counterparts' desire them to do. In this way, a standard of judgement emerges from the armchair, a standard of judgement that we seem to be able to understand in purely descriptive terms. This is the seventh thing we know from the armchair.

Note that this standard of judgement, and its corresponding evaluative and deontic claims, enjoys a certain privilege over all of the other standards of judgement, and their corresponding evaluative and ought claims, that are true of agents in virtue of their being members of other goodness-fixing kinds, goodness-fixing kinds such as *human being*. The privilege lies in the fact that the evaluative and ought claims associated with it are true of any being who can engage in Descartes' thought experiment simply in virtue of the fact that they are beings of that kind. It is therefore a standard of judgement for all such beings, and they are corresponding evaluative and ought claims that are true of all such beings, whether human or non-human. It is a standard of judgement that applies to all rational agents simply in virtue of the fact that they are rational agents, and that means that we can reasonably expect all rational agents with the requisite capacities to conform to them.

It must be admitted that the ought-claims we have derived are quite modest. For example, they leave it open whether we have reasons to do what we morally ought to do. But it should be clear that the heavy-lifting required to address that further question has already been done, for in order to answer it we need to focus on our conception of the ideal agent and see whether every agent's ideal counterpart desires that they act in recognizable moral ways. The question, in other words, is whether it follows from the fact that agents have and robustly exercise maximal capacities to realize their intrinsic desires, to have knowledge of the world in which they live, to be instrumentally rational, and to will, that they care about recognizably moral ends. My own view is that it does, but the argument for this conclusion will have to be given on another occasion.

REFERENCES

- Aristotle 1984: *Nicomachean Ethics*, W.D. Ross (trans.), revised by J.O. Urmson, in *The Complete Works of Aristotle*, The Revised Oxford Translation, vol. 2, Jonathan Barnes (ed.), Princeton: Princeton University Press, 1984.

Bratman, Michael 1987: *Intentions, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press)

Danto, Arthur 1965: 'Basic Actions' in *American Philosophical Quarterly* (2), pp. 141-148

Davidson, Donald 1963: 'Actions, Reasons and Causes' reprinted in his *Essays on Actions and Events*, (Oxford: Oxford University Press, 1980) pp.3-20.

Davidson, Donald 1971a: 'Agency' reprinted in his *Essays on Actions and Events* (Oxford: Oxford University Press, 1980) pp.43-6

Davidson, Donald 1971b: 'Psychology as Philosophy' reprinted in his *Essays on Actions and Events* (Oxford: Oxford University Press, 1980) pp.229-245

Davidson, Donald 1973: 'Freedom to Act' reprinted in his *Essays on Actions and Events* (Oxford: Oxford University Press, 1980) pp.63-83.

Davidson, Donald 1976: 'Hempel on Explaining Action' reprinted in his *Essays on*

Actions and Events (Oxford: Oxford University Press, 1980) pp. 261-277

Descartes, René 1637-1642: *Discourse on Method and the Meditations* (Penguin, 1984).

Foot, Philippa 2001: *Natural Goodness* (Oxford: Oxford University Press).

Hempel 1961: 'Rational Action' reprinted in Norman S. Care and Charles Landesman (eds) *Readings in the Theory of Action* (Bloomington, IN: Indiana University Press, 1968), pp. 285-6.

Johnston, Mark 2010: *Surviving Death* (Princeton: Princeton University Press).

Peacocke, Christopher 1979: *Holistic Explanation: Action, Space, Interpretation* (Oxford: Oxford University Press).

Price, Huw 1996: *Time's Arrow and Archimedes Point* (New York: Oxford University Press).

Sehon, Scott 2005: *Teleological Realism: Mind, Agency, and Explanation* (Cambridge, MA: MIT Press)

Thomson, Judith Jarvis 2008: *Normativity* (Chicago: Open Court Publishing Company).

Williams, Bernard, 1978: *Descartes, The Project of Pure Inquiry* (London: Penguin)