Chapter 4
Beyond Belief and Desire: or, How to Be Orthonomous

Michael Smith

Abstract The standard belief-desire account of the explanation of action is inadequate to the task of explaining even the very simplest actions. We must suppose instead that three psychological states are in play when we explain action, not just two: desire, belief, and the exercise of the capacity to be instrumentally rational. Once we enrich our understanding of action explanation to acknowledge the causal role played by an agent's exercise of his rational capacities, much richer accounts of action explanation come into view, accounts that highlight the many different ways in which agents' actions can be explained by their rational capacities. Of special interest are cases in which agents' actions are explained by their failure to exercise their rational capacities, where these are capacities that they possess, and cases in which their actions are explained by their failure to exercise their rational capacities, where these are capacities that they do not possess. Richer accounts of action explanation such as these suggest a distinctive story about the conditions under which people are responsible for wrongdoing, a story with surprising implications for our understanding of what it is for an agent's moral beliefs to be justified.

4.1 Introduction

The standard belief-desire account of the explanation of action, at least in the form in which that account was put forward by Donald Davidson, is inadequate to the task of explaining even the very simplest actions (Davidson 1963). If some version of the standard account is correct, then we must suppose that it is a variation on the version put forward by Carl G. Hempel (1961) prior to Davidson. According to Hempel, three psychological states are in play when we explain action, not just two as Davidson supposes. Desire and belief are part of the explanation of every action, but so too is the capacity to be instrumentally rational, a capacity that is but one among many capacities rational agents possess (Smith 2009).

Once we enrich our understanding of action explanation to acknowledge the causal role played by an agent's exercise of his rational capacities, much richer accounts of action explanation come into view, accounts that highlight the many

M. Smith (✉)
Department of Philosophy, 1879 Hall, Princeton University, Princeton, NJ 08544, USA
e-mail: msmith@princeton.edu

N.A Vincent et al. (eds.), *Moral Responsibility*, Library of Ethics and Applied Philosophy 27, DOI 10.1007/978-94-007-1878-4_4,

different ways in which agents' actions can be explained by their rational capacities. Of special interest are cases in which agents' actions are explained by their failure to exercise their rational capacities, where these are capacities that they possess, and cases in which their actions are explained by their failure to exercise their rational capacities, where these are capacities that they do not possess (Smith 2003). Richer accounts of action explanation such as these suggest a distinctive story about the conditions under which people are responsible for wrongdoing, a story with surprising implications for our understanding of what it is for an agent's moral beliefs to be justified.

In the first section of the paper I explain why, and in what way, we need to go beyond the standard belief-desire account of action explanation. In the second section I outline the much richer kinds of explanation of action that come into view once we go beyond the standard account in the way suggested, and I describe the conditions of moral responsibility that they suggest. I also provide further support for this story about the conditions of responsibility by bringing out the similarities between it and the conception of responsibility found in the criminal law. And then in the third and final section I outline some of the surprising implications of this conception of responsibility.

4.2 Beyond the Standard Belief-Desire Account of the Explanation of Action

Consider a very simple case of action. Suppose that John non-instrumentally desires more than anything else to get muscly and believes with as much certainty as he believes anything that he can get muscly by exercising. Does it follow that he will exercise, if he does anything at all intentionally?

According to a principle that Donald Davidson accepts, this does follow. The principle is this:

> P1: If an agent wants to do x more than he wants to do y and believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally. (Davidson 1970)

In our example, John satisfies this condition, so by Davidson's lights it follows that he will exercise if he does anything intentionally at all. I take it that this is why Davidson thinks that just two psychological states figure in the explanation of action. All we need to know in order to know what John will do intentionally, if he does anything intentionally, is what he desires and believes, or so Davidson seems to think. However a moment's reflection makes it clear that this isn't so.

Given P1, what should we say about the case in which an agent wants x exactly as much as he wants y, but wants each of these more than he wants anything else, and believes himself free to do various things, including either x or y? If P1 tells us everything we need to know, then all we could say is that he will do either x or y, if he does anything intentionally. But as we know from reflection on Buridan's Ass type cases, this isn't all that we can say. When a rational agent goes to the

supermarket and is confronted by three identical boxes of cereal, desires most to take one or another box, but desires to take each box exactly as much as he desires to take the others, he may still take (say) the one in the middle intentionally. This is because rational agents possess the capacity to just *pick* an alternative when their antecedent desires and beliefs leave them indifferent. They choose one intentionally despite the fact that they don't antecedently desire it more (Raz 1999:100).

To tell us everything we need to know in order to know what an agent will do intentionally, if he does anything intentionally, P1 would therefore need to be supplemented with an account of the role of this distinctive capacity that rational agents possess to pick or chose, a capacity whose exercise explains why they are not flummoxed, suffering the counterpart of starving to death in Buridan's Ass cases. This is therefore a psychological state of great normative significance, and it is also one whose exercise turns out to be empirically tractable. According to one study, for example, when the choice is between three or four or five identical items, as in the case of choosing a box of cereal in a supermarket, rational agents tend to avoid the endpoints, opting for the item in the middle (Christenfeld 1995).

It might be objected that this is all confused. If rational agents in supermarkets intentionally choose items in the middle of a row of identical items, then it follows that they must have at least some non-instrumental desire for things in the middle, a non-instrumental desire that breaks the alleged tie among agents' non-instrumental desires for objects in middle, those on the right, and those on the left. It might be thought that this follows from what it is to desire something. If a desire is just a disposition to choose, then there is no conceptual space for the idea of an additional capacity to choose. The agents in question might not desire the one in the middle more *antecedently*, but they do when they act.

But a little reflection suggests that this is not really an objection to what's been proposed. If the so-called desire for things in the middle only manifests itself in circumstances in which a tie needs to be broken between alternatives that can't be discriminated between by an agent's other non-instrumental desires – if it is the non-instrumental desire, when an agent's other non-instrumental desires leave a choice underdetermined, for those things in the middle, and if it is constitutive of being rational that agents have some such non-instrumental desire when a tie needs to be broken – then there is only a verbal difference between the suggestion that agents have such desires, and the suggestion that they have the capacity to choose an alternative when their desires leave them indifferent between alternatives, a capacity that they tend to exercise by going for the thing in the middle. The reply to the objection thus concedes everything that is at issue.

Once we allow that an additional role may be played in intentional action by the exercise of an agent's rational capacities, as in Buridan's Ass cases, the question immediately arises whether there are other situations in which a role is played by an agent's exercise of his rational capacities. And the answer is that there are. Almost ten years prior to Davidson, Hempel had put forward his own version of the standard account of action explanation. In the course of doing so, he had pointed out that whenever an agent acts on his desires and beliefs, he must also exercise a distinctive rational capacity to put his desires and beliefs together.

Consider once again our example. Suppose that John does non-instrumentally desire to get muscly more than he desires anything else and believes that he can get muscly by exercising with more certainty than he believes anything else. If he is instrumentally irrational he will not form the instrumental desire to exercise, and, absent the formation of that instrumental desire, he won't exercise (Hempel 1961:266–67). For desire and belief even to begin to play the role that Davidson describes in P1, we therefore need to suppose that a ubiquitous role is played by yet another psychological state. Hempel himself calls this state the agent's being rational, but in fact the psychological state in question is both more specific than this, and we have to understand the role that it plays in a certain way.

If an agent with non-instrumental desires and beliefs is to act at all, he must *have and exercise* the capacity to be *instrumentally rational*. It would not be enough for him merely to have the capacity to be rational, where this is a capacity that he may or may not exercise. To return to our example, even if John does have the capacity to be instrumentally rational, if he does not exercise it on the occasion, then he still will not form the instrumental desire to exercise, and, absent the formation of that desire, he will not exercise. Moreover, not just any old rational capacity will do the job. It wouldn't be enough if the agent exercised his capacity to form his beliefs in the light of the available evidence, for example. Indeed, the exercise of that capacity isn't even necessary for an agent to act. Having means-end beliefs is enough. How he came by his means-end beliefs is neither here nor there.

To see more precisely what the distinctive causal role is that's played by an agent's possession and exercise of his capacity to be instrumentally rational, we need to consider a slightly less simple case of action explanation, a variation on the case that we have discussed thus far. Suppose that John has a non-instrumental desire to get muscly and that he believes there are two ways in which he could do so. He believes that he could get muscly by exercising a lot, and he also believes that he could get muscly by exercising less, but taking pills as well. If he does it by exercising a lot, then it will take longer to get muscly, whereas if he does it by using the combination strategy, then he will get muscly sooner, but once the musculature is achieved, it will last equally long either way.

If John is as confident about one of these causal claims as he is about the other – equally confident that exercising a lot will cause him to get muscly and that exercising less and taking pills will cause him to get muscly – then, assuming that he doesn't care whether he gets muscly sooner or later, it follows that, if he were fully instrumentally rational, he would be indifferent between these options. His instrumental desire to exercise a lot and his instrumental desire to exercise less and take pills would be equally strong. He would be in a Buridan's Ass situation, and would therefore need to just pick an option.

But now suppose that John is equally confident about both strategies and that he opts for the combination strategy. If John's possession and exercise of the capacity to be fully instrumentally rational is part of the explanation of his pursuit of the combination strategy, then we already know something about what he would have done if the option of exercising and taking pills hadn't been available to him. John

would have exercised, notwithstanding the fact that it would take him longer to get muscly because, being fully instrumentally rational, he has an instrumental desire just to exercise waiting in the wings to produce action should it turn out that the combination strategy isn't available.

If, however, John is less than fully instrumentally rational – if, say, he has a tendency not to form instrumental desires when gratification is significantly delayed – then we have no grounds for supposing that he would have exercised if the option of taking pills hadn't been available to him. For though it follows from the fact that he exercises and takes pills intentionally that he is at least somewhat instrumentally rational, there is no reason at all to suppose that he is sufficiently instrumentally rational to have formed the desire simply to exercise as well and have it on standby. Indeed, if he has a tendency not to form instrumental desires when gratification is significantly delayed, there is every reason to suppose that he isn't sufficiently instrumentally rational to have formed that desire. Which counterfactuals are true of John thus depends on which explanatory hypothesis is correct.

Here, then, are the questions we must ask. Does John exercise and take pills because he is fully instrumentally rational and picks? Or does he exercise and take pills because, though he is less than fully instrumentally rational, since he had the option of exercising and taking pills, he didn't have to delay gratification? This is an empirical question, one whose answer is fixed by whatever psychological states are the causal antecedents of John's action. If John's action is caused by his possession and exercise of the capacity to be fully instrumentally rational, then he would have just exercised if the combination strategy hadn't been available. But if John's action is caused by his possession and exercise of a diminished capacity to be instrumentally rational, then he wouldn't have just exercised if the combination strategy hadn't been available.

Let's sum up the discussion thus far. We have seen that the standard belief-desire account of the explanation of action, at least in the form proposed by Davidson, is inadequate. We must suppose, with Hempel, that agents possess not only desires and beliefs, but also the capacity to be instrumentally rational to some extent. Moreover we must also suppose that their possession and exercise of this capacity plays its own distinctive explanatory role, complementary to the role played by their desires and beliefs, whenever agents act. We must also suppose that agents possess other rational capacities as well, capacities like the capacity to pick an alternative when antecedent desires and beliefs leave them otherwise indifferent.

I take this to be sufficient reason to move to a Hempelian, rather than a Davidsonian, conception of the standard account of action explanation. An agent's actions are explained by psychological states of three kinds, not just two: his desires, his beliefs, and the exercise of his rational capacities. But once we acknowledge that an agent's possession and exercise of rational capacities plays a distinctive role in the explanation of his actions, our eyes are opened to the possibility of much richer accounts of action explanation. These richer accounts in turn suggest a way in which we might begin to flesh out the conditions of responsibility.

4.3 The Nature of Responsibility

Consider once again the case in which John exercises and takes pills because, though he is less than fully instrumentally rational, he didn't have to delay gratification. It turns out that there are two possibilities here, depending on whether we suppose that John has a diminished capacity to be instrumentally rational which he fully exercises, or an undiminished capacity to be fully instrumentally rational that he fails to exercise on the occasion. There are therefore two corresponding further explanations of John's behaviour, depending on which of these possibilities is realized. In the first, John exercises and takes pills because he lacks the capacity to be fully instrumentally rational. In the second, he exercises and takes pills because, though he has that capacity, he fails to exercise it.

What's especially striking about these two further explanations is that they bear their relationship to ascriptions of responsibility more or less on their face. It follows from the very nature of responsibility that an agent who fails to act permissibly because he lacks the rational capacities required to act in that way is not responsible for failing to act permissibly. He is not responsible because his incapacity serves as an excuse. This is why children, those who are deranged, and those with volitional deficiencies like Obsessive Compulsive Disorder (OCD) are so often excused when they act wrongly. Children, the deranged, and those with OCD lack certain rational capacities, so when they act wrongly because they lack these capacities – and note that this needn't be true every time they act wrongly – they are thereby excused. Ascriptions of responsibility for wrongdoing are assignments of fault and these agents are not at fault.

By contrast, an agent who acts impermissibly because he fails to exercise rational capacities that he possesses is responsible for failing to act in that way. He is responsible precisely because he has no excuse. This is why someone who suffers from (say) weakness of will isn't treated like a child, someone who is deranged, or someone with OCD. Those who suffer from weakness of will have the capacity to will otherwise, but fail to exercise it. When they act impermissibly they are therefore liable to be held responsible because they are expected to exercise their capacity to will otherwise. The fact that they don't exercise their capacities is the problem, not an excuse. Fault is properly assigned to them.

Just to be clear, note I do not intend these claims to express a substantive moral view. They are meant to express conceptual claims, or, if you prefer, metaphysical claims, about what it is for an agent to be responsible. It is a priori that an agent is responsible for wrongdoing just in case he acts impermissibly without justification or excuse, and it is similarly a priori that when an agent's wrongdoing is explained by his lacking certain rational capacities, he has an excuse. This is why I said earlier that I took myself to be spelling out the nature of responsibility. These claims spell out *internal* correctness conditions of responsibility ascriptions, not substantive moral commitments.

There is, of course, a substantive moral view according to which we should treat agents in the way we typically treat responsible agents – we should, for example, punish them – only if they are responsible. In the theory of punishment, this is

the view held by retributivists. Others disagree. They hold that we should sometimes treat agents in the way we typically treat responsible agents even when they aren't responsible. In the theory of punishment, this is the view held by utilitarians. They believe that the fact that it would maximize happiness to punish someone (say) is always a good reason to do so, whether he is responsible or not.

Moreover, again just to be clear, the claim that an agent is excused of a wrong that he has done isn't a substantive moral claim either. In particular, it is not, and does not imply, the claim that the agent in question should be left free to do whatever he pleases. To say that an agent is excused of wrongdoing is simply to say that the wrong he did was not his fault. But even when the wrong that someone does is not his fault, his acting wrongly in the circumstances in which he did might still provide others with grounds for coercing him. Those who hold different substantive moral views can and should agree with this.

For example, retributivists and utilitarians can and should agree that someone who does wrong, but who is excused of that wrongdoing because he is deranged, may not be someone who should be left free to do what he pleases. There may be a justification for using coercive means to restrain him if he won't listen to reason. The crucial point is simply that the justification for coercing him could not be that he did something that was his fault. The justification would have to be that (say) what he did, together with his being deranged, shows that he is a danger to himself and others (this is the sort of justification that might be given by those attracted to retributivism on Kantian grounds, though of course this is no part of retributivism itself), or that coercing him would maximize happiness (this is the sort of justification that might be given by a utilitarian).

If what has been said so far is along the right lines, then this suggests a way in which we might proceed in order to fully spell out the conditions of responsibility. We might proceed by coming up with an exhaustive list of the rational capacities whose possession and exercise would be necessary for agents to be responsible when they act impermissibly. We have already seen that at least two such capacities would be required: the capacity to pick an alternative when our non-instrumental desires and means-end beliefs underdetermine the choice between alternatives and the capacity to put our non-instrumental desires together with our means-end beliefs. Are there any others?

A capacity suggested by the foregoing discussion is the capacity to form beliefs in the light of the evidence available to us. Someone who harms another in the course of satisfying some instrumental desire he has is not excused of wrongdoing merely because he had no idea that harm would result. Ignorance is no excuse because we are expected to exercise such capacities as we have to access relevant evidence, and then to form our beliefs on the basis of that evidence. But an agent who was literally *incapable* of forming the belief that harm would result from something that he does would be excused. If he lacks the capacity to access the relevant evidence, or the capacity to form beliefs in the light of that evidence, then the harm that he causes is not his fault. If he has, but simply fails to exercise, these capacities, however, then the harm he causes is his fault.

What about an agent's non-instrumental desires? Are there capacities to form non-instrumental desires that agents must possess if they are to be responsible when they act impermissibly? This is an issue on which there are deep divisions within philosophical psychology. Some theorists line up behind Hume who thinks that the non-instrumental desires that move us to action are "original existences" (Hume 1740). They think that the answer therefore has to be "No". Others line up behind Kant who thinks that we have the capacity to allow only those non-instrumental desires that accord with universal laws of reasons to motivate us (Kant 1786). They think that the answer has to be "Yes". But without addressing the issue that divides these theorists head-on, note that there is at least one reason for supposing that the answer must be "Yes", a reason that can be appreciated by followers of both Hume and Kant alike.

According to the doctrine in meta-ethics known as *Judgement Internalism*, if an agent believes that he ought to φ, then is motivated to φ, at least insofar as he is rational (Smith 1994:63–84). Given that an agent's motivations are constituted by his non-instrumental desires and means-end beliefs, *Judgement Internalism* implies that if an agent believes that his φ-ing is a basic wrong – that is, if he believes that it is a wrong simply in virtue of its being a φ –ing – then, insofar as he is rational, he will have a non-instrumental aversion to φ-ing. The capacity to be rational thus mediates between an agent's beliefs about the things that would be basic wrongs to do and his non-instrumental aversions. For an agent to fail to have a non-instrumental aversion to doing what he believes it would be a basic wrong to do therefore implies irrationality on his behalf, where, as above, this irrationality might therefore be grounded in two quite different sorts of fact about him.

On the one hand, the agent's irrationality might be grounded in an incapacity to acquire non-instrumental aversions that accord with his beliefs about basic wrongs. In commonsense terms, this would be for him to do wrong because he has no capacity for self-control. He knows what he should do, but he can't get himself to want to do it. If this is the form that his irrationality takes, then, as before, if he acts wrongly, he has an excuse, for his inability to control himself constitutes his excuse. His acting wrongly is not his fault. Alternatively, though the agent possesses the capacity to control himself, his irrationality might be grounded in his failure to exercise this capacity. If this is the form that his irrationality takes, then, as before, he has no excuse, for his wrongdoing is his fault.

Moreover, note that there will evidently be complicated mixed cases. If an agent does something wrong because he is (say) crazed on drugs, but he didn't lack the capacity for self-control when he took the drugs, and he knew at that earlier time that taking drugs would cause him to become deranged and do something wrong, then though there is a sense in which he does wrong because he lacks the capacity for self-control – when he was crazed, he couldn't control himself – his doing wrong isn't grounded in his lack of self-control in the way it would have to be to constitute an excuse. This is because his doing wrong can be traced to his earlier failure to exercise the capacity that he possessed for self-control when he took the drugs. His wrongdoing may therefore still be his fault.

So far we have focused on rational capacities that mediate between an agents' beliefs about basic wrongs and their non-instrumental desires. But *Judgement Internalism* implies that additional *cognitive* capacities will also be required if agents are to be responsible for acting impermissibly. We have already seen that when it comes to belief formation, agents who have capacities to access relevant evidence, and then to form their beliefs on the basis of that evidence, are expected to do so. When it comes to exercises of self-control, what is especially important is therefore that agents exercise their capacity to access relevant evidence about which acts are wrong, and that they exercise their capacity to form beliefs on the basis of that evidence. The upshot is thus that even those who do have the capacity for self-control, and who exercise that capacity, may still not be responsible for wrongdoing if they are not responsible for the moral beliefs on which they act.

For example someone who doesn't know that what he is doing is wrong will be excused for his wrong-doing if his ignorance results from an inability to know that what he is doing is wrong: that is to say, he will be excused if he is ignorant because he lacks either the capacity to access the available evidence, or the capacity to form beliefs in the light of that evidence. However if he doesn't know that what he is doing is wrong because he has, but fails to exercise, his capacities to access available evidence, or to form beliefs in the light of that evidence, then he has no excuse. His wrongdoing is his fault because he could and should have exercised his capacity to know what's right and wrong.

Let's step back for a moment. What's remarkable about this account of the internal correctness conditions of ascriptions of responsibility isn't just that it can be derived entirely a priori by reflection on what the rational capacities are whose exercise might play a role when an agent acts, but also that it bears a striking similarity to conceptions of criminal responsibility that we find in the law. This shouldn't really be surprising, given that legal conceptions of criminal responsibility have a distinctively retributivist flavour. But the fact that it is so adds additional credence to the account of responsibility I have been sketching. Here are some examples, just to drive the point home.

The minimum age at which someone can be held criminally responsible in Australia is ten, as younger children are deemed to be incapable of knowing the difference between right and wrong (Australian Government 2005). Australian law also includes the doctrine of *Doli Incapax*. According to this doctrine, though children between the ages of ten and fourteen may possess the capacity to know the difference between right and wrong, they are presumed not to. The presumption therefore has to be proved mistaken before a criminal case can proceed. Both of these ideas fit very smoothly with the idea that agents who are incapable of knowing that what they are doing is wrong, children being a prime example, are excused when they act wrongly.

The law on insanity as a defence in criminal cases builds on a related idea. The law on insanity was developed in *Queen v M'Naghten* in 1843 when Daniel M'Naghten approached a man who he believed to be Sir Robert Peel, the then Prime Minister of England, and shot him in the back, so killing him. When M'Naghten was tried for the man's murder it emerged that he firmly believed that Peel was out to

kill him. After testimony from medical experts, M'Naghten was found not guilty by reason of insanity. The decision caused such a controversy that the House of Lords asked the Lords of Justice to formulate a strict definition of when insanity could be used as a defence against criminal charges. According to the definition, now known as the "M'Naghten Rule", insanity is a defence only if:

1. At the time that the act was committed
2. the defendant was suffering from a defect of reason, from a disease of the mind, which caused
3. the defendant to not know
   a. the nature and quality of the act taken or
   b. that the act was wrong. (Hall 2008:226–27)

The M'Naghten Rule is widely accepted in jurisdictions influenced by English law. In crucial respects, the idea behind the M'Naghten Rule is much the same as before. An agent who has lived for long enough to develop the capacity to know what's right and wrong, but who suffers from some "disease of the mind" that destroys that capacity, is excused of wrongdoing if his wrongdoing is the result of his incapacity to know either what he is doing or that what he is doing is wrong.

Critics of the M'Naghten Rule insist that its exclusive focus on cognitive incapacities results in too narrow a conception of the insanity defence. In some jurisdictions, it is therefore supplemented with what is known as the "Irresistible Impulse Test". The Irresistible Impulse Test, developed in a decision by the Alabama Supreme Court in the USA in *Parsons v State* in 1887, holds that even if an accused party could tell right from wrong, he may still be excused by reason of insanity:

A. if mental disease caused the defendant to so far lose the power to choose between right and wrong and to avoid doing the alleged act that the disease destroyed the defendant's free will, and
B. if the mental disease was the sole cause of the act. (Lippman 2009:279)

The Irresistible Impulse Test thus also fits very smoothly with the account of the conditions of responsibility sketched in the previous section. What's relevant to the Irresistible Impulse Test is the capacity for self-control: that is, an agent's capacity to form non-instrumental aversions to doing those things he believes to be basic wrongs. If an agent lacks the capacity for self-control, and acts because he lacks that capacity, then he is excused because his act is not his fault.

Let's sum up. Not only are actions are always explained, inter alia, by agents' exercise of their capacity to be instrumentally rational, a capacity they might possess to a greater or a lesser extent, but many actions are also explained by agents' other rational capacities, sometimes by their exercise of these capacities, and sometimes by their failure to exercise them. These other rational capacities include the capacity to access available evidence; the capacity to form beliefs in the light of that evidence, both beliefs about means to ends and also beliefs about what's right and wrong; and

the capacity to exercise self-control, that is, the capacity to form non-instrumental desires and aversions in the light of beliefs about which acts are right and wrong.

Cases in which actions are explained by agents' failure to exercise their rational capacities are in turn of two types. There are those in which their actions are explained by a failure to exercise rational capacities that they possess, and there are those in which their actions are explained by the fact that they do not possess those capacities in the first place. This is important when it comes to ascriptions of responsibility, because it is part of the internal correctness conditions of such ascriptions that when agents act wrongly because they lack some relevant rational capacity, they are excused, whereas when they act wrongly because they have, but fail to exercise, some relevant rational capacity, they are not excused. Distinctions widely made within the criminal law give some support to these ideas.

4.4 Implications

At the very beginning I said that we need to move beyond Davidson's version of the standard story of action explanation, and that, when we do, a distinctive story emerges about the conditions under which people are responsible for wrongdoing, a story with surprising implications about the justification of an agent's moral beliefs. Let me now spell out some of these implications.

In "Sanity and the Metaphysics of Responsibility", Susan Wolf describes an agent whose responsibility is seriously in doubt.

> JoJo is the favourite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not coerced to do these things, he acts according to his own desires. Moreover these are desires he wholly wants to have. When he steps back and asks 'Do I really want to be this sort of person?' his answer is resoundingly 'Yes,' for this way of life expresses a crazy sort of power that forms part of his deepest ideal. . .In light of JoJo's heritage and upbringing — both of which he was powerless to control — it is dubious at best that he should be regarded as responsible for what he does. It is unclear whether anyone with a childhood such as his could have developed into anything but the twisted and perverse person that he has become. (Wolf 1987:53–54)

Wolf doesn't explicitly describe JoJo in the detailed terms that we have seen are crucial for understanding his responsibility. It is, however, easy to see how different ways of filling in the details of his story would affect his responsibility, and while certain of these ways of filling the story deliver unsurprising results, others deliver results that are much more surprising.

Here is one way in which JoJo's story might be filled in. JoJo's father, Jo the First, developed a talent for the manipulation of other people for his own purposes when he was a young boy. It was this talent for manipulation that enabled him to become ruler of the small, undeveloped country, in which he grew up. Once he became

leader, Jo the First saw to it, sometimes by manipulation, but when manipulation failed, by whatever means were necessary, that all of those who lived in the country did exactly what he wanted. No one spoke ill of him, not even in private, for fear of dire consequences.

When JoJo was born, Jo the First fixated on him. He saw in JoJo someone who could see to it that no one would speak ill of him even after he died. He therefore gave JoJo a special education which consisted of humiliating him and then building him back up by making him believe that the only way he could have any worth at all was by getting his father's approval, something that he could do by emulating his father's behaviour, singing his praises, and generally seeing to it that others did nothing that his father wouldn't like. When people criticized his decision to home-school his son, he had them silenced. JoJo too therefore developed a talent for the manipulation of other people for his own purposes – many of which were of course Jo the First's purposes – when he was a young boy.

After Jo the First died, JoJo took over as ruler of the country, doing many of the same sorts of things his father had done, including sending people to prison or to death or to torture chambers on the basis of whim. He did all of this willingly, constantly singing the praises of his father and seeing to it that no one ever said anything to challenge the official view of his father as a great man. Since he was following in his father's footsteps, this meant that he had to see to it that people treated him as a great man as well. When he had a son, he saw in him exactly what his father had seen in him, and decided to give him the same education that he had received. When people criticized his decision to homeschool his son, he had them silenced. Brutally.

If we fill in the details of JoJo's story in this way, then Wolf is surely right that his responsibility is seriously in doubt, as JoJo seems to have been brainwashed to do his father's bidding. His belief that his father is a great man, the belief which sustains his desire to emulate his father's actions, is kept in place not by his assessment of the evidence available to him for his father's greatness, but rather by his need for his father's approval. JoJo seems to lack the capacity to form beliefs about whether his father's, and hence his own, actions are right or wrong in the light of the evidence available to him. He therefore seems not to be responsible for reasons similar to the reasons why children and the insane are not responsible.

Wolf herself points out that JoJo is similar to those who fall under the M'Naghten Rule (Wolf 1987:55). Indeed, she suggests that we should suppose that JoJo is "insane" in an "admittedly specialized sense":

> [A]lthough like us, JoJo's actions flow from desires that flow from his deep self, unlike us, JoJo's deep self is itself insane. Sanity, remember, involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability. (Wolf 1987:56)

There are, of course, differences between JoJo and those who are criminally insane. To return to the M'Naghten Rule, JoJo's brainwashing doesn't seem to have caused a "disease of the mind", unless, of course, we are using the term "disease"

in a highly metaphorical sense. (I note in passing that the M'Naghten Rule itself doesn't define what a disease of the mind is, and that in some jurisdictions it is interpreted very broadly.)

On the other hand, however, JoJo's brainwashing plainly has caused him to have a mental problem, a problem whose upshot is a lot like the upshot of M'Naghten's disease of the mind. M'Naghten's disease of the mind caused him to be insensitive to evidence as to whether or not Peel was out to kill him in his formation of his belief that this is what Peel was out to do. JoJo's belief that his father's actions are not wrong is insensitive to evidence as to whether or not his actions are right or wrong because his brainwashing has caused his desire for his father's approval to sustain his belief independently of evidence. Perhaps this similarity is all Wolf needs to be right that JoJo is insane "in a specialized sense". We will return to this point below.

It might be helpful if we think in terms of diagrammatic representations. The responsible agent's actions can be represented as in Fig. 4.1.

In Fig. 4.1, the "→"s represent either a relation that is knowledge conducive (1, 2), or the exercise of some relevant rational capacity (3, 4, 5, 6), or causation of a kind that sustains differential explanation (7, 8) (for more on this, see Smith 2004, 2009). In these terms, what's crucial about JoJo, at least when we fill in the details of his story as we did above, is that he lacks a crucial rational capacity that the responsible agent possesses and exercises: specifically, he lacks the capacity at junction 4. This suffices to excuse him from wrongdoing when his wrongdoing is explained by his lack of this capacity. He is excused, in such cases, because his wrongdoing is not his fault.

There is, however, the following rather different way in which the details of JoJo's story might be filled in. As he grew up, JoJo was given a special education, both formal and informal, by his father, Jo the First. Jo the First was brutal, but also articulate and larger than life, much like one of the main characters in a Quentin

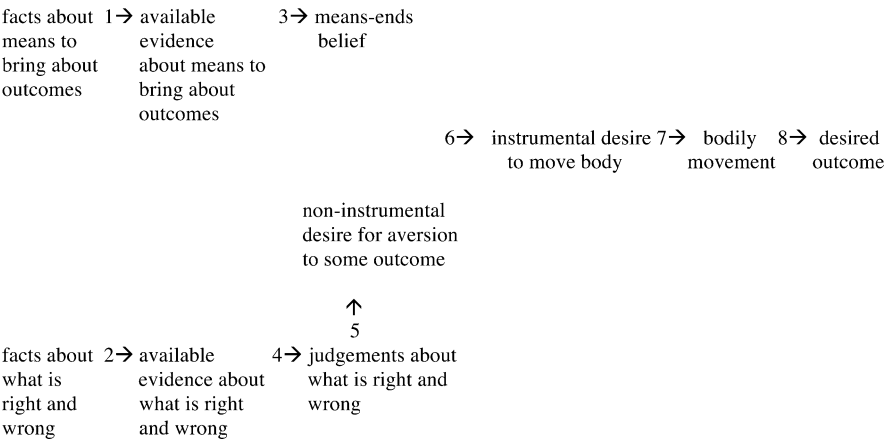

Fig. 4.1 The responsible agent

Tarantino movie. As a result, JoJo came to have the rather idiosyncratic belief that someone who has the sort of power his father has, exercised in the stylish way in which he exercised it, thereby has the right to do whatever he has to do in order to maintain that power, no matter what the consequences are for those he doesn't really care about. JoJo read widely, something his father encouraged, so he realised that his view about his father's entitlements was idiosyncratic. Though this was initially a cause of some cognitive dissonance, one day JoJo came across some books by a German philosopher that contained an elaborate statement and defence of ideas that imply that something like what he had come to believe about his father was in fact true. After reading the German philosopher's work for himself, he concluded that, idiosyncratic though his ideas were, there was also something deeply intuitive about them.

JoJo's interest in philosophy led him to apply to graduate school at an Ivy League university in the USA where the world's leading expert on the German philosopher taught. JoJo's admission file was so strong that he was given a special scholarship. He eventually wrote a dissertation defending his own unusual moral views, drawing on their similarities to those of the German philosopher's, via a wide reflective equilibrium argument, a dissertation that eventually became a celebrated monograph. JoJo's monograph was much discussed by academic philosophers, and also much discussed in the pages of the New York Review of Books and on NPR.

While many reviewers thought that the arguments JoJo gave were spot on, others thought that the position itself, though internally consistent, and though consistent with everything else JoJo believed, depended on basic premises that were themselves manifestly implausible. JoJo's conclusions could be true only if some of the things that they themselves believed were true were false, but they did not think that anything JoJo said had provided them with sufficient reason to reject the truth of what they believed. Some of these reviewers, ignorant of JoJo's history, went so far as to say that it just as well that JoJo was a harmless academic, as by this time he had secured a professorship of his own at a prestigious left-wing university in northern California.

JoJo saw nothing unusual in the fact that his colleagues had such starkly different opinions about his work. He had come to the view years earlier that there are no philosophical theories on any subject matter that command universal assent. As he saw things, all philosophical theories are defended via wide reflective equilibrium arguments of the kind he had given, and this meant that the very deepest philosophical disagreements amounted to disagreements about fundamental premises: that is, they were disagreements about which beliefs are, and which are not, supposed to survive such an argument. He went on to become something of an academic celebrity, universally admired for his charm and wit and intelligence and loyalty to his graduate students and close colleagues, but also feared by those who experienced how ruthlessly dismissive he could be of those with whom he saw no profit to engage.

After Jo the First died, JoJo seized the opportunity to put his ideas into practice on a much larger scale. He returned home to take over as ruler of his country. He did many of the same sorts of things his father had done, including sending people

to prison or to death or to torture chambers on the basis of whim. He did all of this willingly, constantly singing the praises of his father and seeing to it that no one ever got away with challenging the official view of his father as a great man. When people criticized his father, or him, he would send them a copy of his monograph and a long list of references to papers in academic journals in which philosophers wrote at length defending the essentials of his views. When he eventually had a son, and people criticized his decision to homeschool him, he would remind them of that one of their heroes, John Stuart Mill, was himself homeschooled by his father. If they persisted with their criticisms, and became disruptive or unpleasant, he had them silenced. Brutally, but stylishly.

If we tell JoJo's story in this way, is it plausible to suppose that he is not responsible for his wrongdoing? By contrast with the earlier telling of his story, it doesn't seem that JoJo's beliefs are the product of brainwashing or wish fulfilment. They are rather the product of deep thought and rational assessment. Indeed, when we tell the story in this way, JoJo *seems* to be at least as rational as anyone we are likely to meet, more rational than most. Moreover he seems to be exceptionally diligent in his exercise of his rational capacities. Given his education and his dedication to the academic enterprise, we might even be tempted to suppose that JoJo's moral beliefs, though false, are as justified as anyone's could be. JoJo thus doesn't seem to lack any rational capacities. But if he lacks no rational capacities, then how could he not be responsible?

What this way of telling JoJo's story teaches us, I think, is that we need to get much clearer about what's happening at both junctions 2 and 4 in Fig. 4.1. In order to do this, it will be helpful if we first of all think about junctions 1 and 3. Imagine someone who lacks peripheral vision, and to whom it therefore seems that there are no objects in his immediate environment when in fact there are. He therefore regularly acquires beliefs about things he can do that are false: for example, he regularly acquires the belief that it is safe to cross the road, when in fact crossing the road would cause him to be hit by a car. Is such a person responsible for his false beliefs? This question would not be easy to answer in practice, but we know how to answer it in theory.

Assuming that the person we are imagining didn't cause his own lack of peripheral vision, he certainly isn't responsible for its seeming to him that there are no dangerous objects in his environment when there are, because he can't help how things seem to him. That is just a given, a function of his perceptual system. Of course, since experience has presumably taught him that he shouldn't trust how things look to him in forming his beliefs about how things are, he may well be responsible for not pausing to ask whether things really are, in every detail, the way they look to him to be. But it is an empirical question whether human beings really do have the capacity to resist the natural inclination to form perceptual beliefs on the basis of perceptual appearances during the hustle and bustle of daily life. Perhaps the connection between perceptual appearances and perceptual beliefs at such times is so immediate that that kind of second-guessing simply isn't realistic. Either human beings don't have such a capacity, or, if they do, it is a capacity that it would be very difficult for them to exercise.

If this is right, then the person we are imagining may not be responsible for his false beliefs at all, or his responsibility might be seriously mitigated. He is not responsible if he is incapable of having the world seem to him to be the way it really is and he non-culpably finds himself unable to resist the natural tendency to believe that things are how they seem to him to be on the basis of the indirect evidence available to him. His responsibility is mitigated if, though he has the latter capacity, it would be very difficult for him to exercise it during (say) the hustle and bustle of daily life. We might put the same points more explicitly in terms of the language of Fig. 4.1 as follows. Focus on the case in which he isn't responsible at all. The person who lacks peripheral vision has two problems. First, perceptual evidence about how things are in certain regions of his immediate environment isn't available to him because things don't seem to him to be the way that they are. And second, indirect evidence about how things are in those regions – for example, indirect evidence that, for all he knows, there is something in those regions of his immediate environment – though available to him, isn't evidence to which he has the capacity to be sensitive during the hustle and bustle of daily life. He therefore isn't responsible for (say) his false belief that it is safe to cross the road because his false belief isn't his fault.

Let's now consider what to say about JoJo in the light of this. Focus on junctions 2 and 4. JoJo acquired the false belief that Jo the First was within his rights to do the brutal things he did in the way in which children usually acquire their moral beliefs, that is, by an informal process of socialization. When he became an adult, however, he questioned whether his beliefs were true, and he concluded that they were. He reached this conclusion in two ways. It both seemed to him that they were true – that is, what he believed was, he thought, deeply intuitive – and, furthermore, after thinking long and hard about questions in moral philosophy, he came up with a reflective equilibrium argument for a theory that entailed the truth of the things that he believed. So is JoJo responsible for his false belief? The answer to this question bears certain similarities to the answer we just gave about the person who lacks peripheral vision.

JoJo also has two sorts of problems. First, direct evidence of his father's wrongdoing isn't available to him, because the things that seem permissible to him aren't permissible. His father seems to JoJo to have the right to brutalize people, when in fact he has no such right. This is strictly analogous to what we said about the person who lacks peripheral vision. Second – and this is a difference between JoJo's case and that of the person who lacks peripheral vision – indirect evidence of his father's wrongdoing isn't available to him either. For in order to access indirect evidence of the wrongness of his father's actions, JoJo would have to be able to construct a theory that entailed that his father's acts were wrong via an attempt to get his beliefs into a wide reflective equilibrium. But he can't. When he succeeds in his attempt to get his beliefs into a wide reflective equilibrium, the theory that he comes up with entails that his father's acts are not wrong.

We can put the same point more simply as follows. For JoJo to be able to access evidence of his father's wrongdoing, there would have to be something that he believes, or something that he feels, or some way that things seem to him to be, that doesn't square with his father's having a right to brutalize people. Absent some

such psychological hook, JoJo will be unable to reason himself to the conclusion that his belief that his father has a right to brutalize people is false because there will be nothing for him to reason from. But there are no such psychological hooks in JoJo. His beliefs, his feelings, the ways things seem to him to be, all of these things square with his belief that his father has a right to brutalize people. The upshot is that JoJo isn't responsible for his false belief that his father has a right to brutalize people. He isn't responsible because he cannot access evidence to the contrary. His false belief isn't his fault.

I said at the beginning that the distinctive story we've told about the conditions under which people are responsible for wrongdoing has surprising implications for the justification of an agent's moral beliefs. These implications are implicit in the conclusions we have just drawn from the second way of filling in the details of JoJo's story. As we have seen, even though JoJo succeeds in getting his beliefs into a wide reflective equilibrium, his moral beliefs are false. Should we suppose that his moral beliefs are justified? There are two ways we could go in answering this question. On the one had, we might suppose that the description of the wide reflective equilibrium procedure itself just is an account of the conditions under which an agent's moral beliefs are justified, so that the answer has to be that JoJo's moral beliefs are justified. On the other, we might wonder whether basic moral beliefs that an agent holds only because he is irrational could ever be justified. Since this seems to be so in JoJo's case, we might conclude that the answer has to be that his basic moral beliefs are not justified. I won't decide between these two ways in which we might answer the question in what follows. I will simply spell out the second way of answering the question in a little more detail.

Think again about the difference between the person who lacks peripheral vision and JoJo. In both cases, the world seems to them to be a certain way when it isn't that way. But in the case of the person who lacks peripheral vision, this fact about him doesn't suggest irrationality of any kind. The defect lies in his perceptual system, not in the capacities he possesses insofar as he is a reasoner. In JoJo's case, by contrast, the fact that the world seem to him to be a certain way when it isn't that way does suggest irrationality of some kind. It suggests irrationality because it entails a limitation on his abilities as a reasoner. There is, of course, an assumption that I'm making here, namely, that knowledge of basic moral truths is a priori accessible. But if this assumption is correct, as I think it is (Smith 1994, Chapters 5 and 6), then given that the fact that the world seems to JoJo to be a certain way in basic moral respects when it isn't that way is what explains his inability to know basic moral truths, it follows that that fact is also what explains why he isn't an ideal reasoner. An ideal reasoner is, after all, someone with the ability to know a priori truths.

If this way of thinking about the justification of JoJo's beliefs is correct, then it follows that we need to radically rethink the epistemic significance of the reflective equilibrium procedure (compare Scanlon 2002). Though JoJo succeeds in getting his beliefs into a wide reflective equilibrium, given that he achieves that wide reflective equilibrium only because he isn't an ideal reasoner, we should conclude that the beliefs he comes up with are not justified. To be justified, an agent's moral beliefs mustn't just be such that they would survive his attempts to get his beliefs into a wide reflective equilibrium. That wide reflective equilibrium itself mustn't be sustained

by the agent's inability to know certain a priori truths. We are therefore led to the conclusion, which may well come as a surprise to some, that whether or not we think that an agent's moral beliefs are justified will depend crucially on what we take the moral truth to be, as this will in turn affect which reasoning capacities we take the justification of an agent's moral beliefs to depend upon.[1]

References

Australian Government. 2005. "The Age of Criminal Responsibility, Australian Institute of Criminology." Accessed 21 January 2011. http://www.aic.gov.au/publications/current%20series/cfi/101-120/cfi106.aspx.

Christenfeld, N. 1995. "Choices from Identical Options." *Psychological Science* 6:50–55.

Davidson, Donald. 1963. "Actions, Reasons and Causes." Reprinted in *Essays on Actions and Events*, edited by Donal Davidson, 3–20.Oxford: Oxford University Press, 1980.

Davidson, Donald. 1970. "How Is Weakness of the Will Possible?" Reprinted in *Essays on Actions and Events*, edited by Donal Davidson, 21–42. Oxford: Oxford University Press, 1980.

Hall, Daniel. 2008. *Criminal Law and Procedure*. New York, NY: Cengage.

Hempel, Carl G. 1961. "Rational Action." Reprinted in *Readings in the Theory of Action*, edited by Norman S. Care and Charles Landesman, 285–86. Bloomington, IN: Indiana University Press, 1968.

Hume, David. 1740. *A Treatise of Human Nature*. Oxford: Clarendon Press, 1968.

Kant, Immanuel. 1786. *Groundwork of the Metaphysics of Morals*. London: Hutchinson and Company, 1948.

Lippman, Matthew. 2009. *Contemporary Criminal Law: Concepts, Cases, and Controversies*. Thousand Oaks, CA: Sage Publications.

Raz, Joseph. 1999. "Explaining Normativity: Reason and the Will." In *Engaging Reason: On the Theory of Value and Action*, 90–117. Oxford: Oxford University Press.

Scanlon, Thomas M. 2002. "Rawls on Justification." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman, 139–67. New York, NY: Cambridge University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Wiley-Blackwell.

Smith, Michael. 2003. "Rational Capacities." In *Weakness of Will and Varieties of Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 17–38. Oxford: Oxford University Press.

Smith, Michael. 2004. "The Structure of Orthonomy." In *Action and Agency*, edited by John Hyman and Helen Steward, 165–93. Cambridge: Cambridge University Press.

Smith, Michael. 2009. "The Explanatory Role of Being Rational." In *Reasons for Action*, edited by David Sobel and Steven Wall, 58–80. New York, NY: Cambridge University Press.

Wolf, Susan. 1987. "Sanity and the Metaphysics of Responsibility." In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by Ferdinand Schoeman, 46–62. New York, NY: Cambridge University Press.

[1] Earlier versions of this paper were presented *at Moral Responsibility: Neuroscience, Organization, and Engineering*, a conference held at Technical University Delft in August 2009; at *Workshop on Reasons and Rational Choice* held at the London School of Economics in January 2011; and at the Philosophy Departments at Lingnan University and Monash University in March 2011. Thanks are due to the many people who gave me comments on these occasions. Special thanks are owed to Jay Wallace for the conversations we had while I was writing up the penultimate version and to Nicole Vincent for her written comments. Work on this paper was completed while I visiting the Humboldt University in Berlin, enjoying the benefits of a Forschungspreis from the Alexander von Humboldt Foundation.