almost certainly did not, as the courts noted, which is why the Law Lords' principle did not lead to their acquittal) their belief was reasonable.[13]

13 I discussed some of these issues in 'A Probabilistic Approach to Moral Responsibility' in Ruth Barcan Marcus *et al* (eds), *Proceedings of the Seventh International Congress of Logic, Methodology, and Philosophy of Science* (1986) 351–66. I am indebted to the many discussions the earlier paper prompted, and must mention discussions with Philip Pettit and Stephanie Lewis.

3 Irresistible Impulse

MICHAEL SMITH

According to the McNaghten Rules of 1843, those charged with a criminal offence may be found not guilty by reason of insanity if they are so affected by a mental disease or defect that, at the time of the offence, they are either unable to understand the criminality of their act or are unable to conform their behaviour to the law. The precise interpretation of the rules is, however, a matter of some legitimate dispute.

In Australian criminal law the McNaghten Rules have been given a fairly wide interpretation. In *R v Porter* Justice Dixon said:

> If through the disordered condition of his mind [the accused] could not reason about the matter with a moderate degree of sense and composure it may be said that he could not know that what he was doing was wrong.[1]

Thus, in Dixon's view, people suffering from delusions, from brainwashing, and even those suffering from emotions and anxiety, should be excused to the extent that their delusory or brainwashed or discombobulated beliefs played a causal role in their conduct.

However some have argued that even Dixon's wide interpretation of the McNaghten Rules is unnecessarily narrow.[2] Dixon's gloss focuses exclusively on *cognitive* impairments, but there would seem to be *non-cognitive* impairments that impact on conduct as well. For example, those who suffer from an irresistible impulse, lacking all self-control, may know perfectly well that what they do is wrong. But if they are literally unable to translate their beliefs into action then, the suggestion goes, their conduct should be excused too. It should be excused because they are just as incapable of conforming their behaviour to the law as those whose reasoning capacities are impaired. Moreover, if there are people who, while not lacking the capacity for self-control entirely, have a capacity that

1 *R v Porter* (1933) 55 CLR 182, 189–190. Dixon J's judgment is quoted in Herbert Fingarette, *The Meaning of Criminal Insanity* (1972) 204 fn 17.

2 C L Ten, *Crime, Guilt and Punishment* (1987) 125.

is none the less diminished or limited in some way, then their liability too should be correspondingly limited.

There is no denying that this even wider interpretation of the McNaghten Rules accords with commonsense. But the mere fact that it accords with commonsense is no guarantee that the idea of an irresistible impulse, and the correlative idea of an agent's lacking all capacity for self-control, or having a diminished capacity for self-control, can be spelled out in a coherent and plausible way. My aim in the present essay is to focus attention on whether or not this can be done. To anticipate, I will argue that though these ideas make good sense, once we understand why they make good sense we also see some reason to suppose that they have only a limited application as excuses.

The Standard Humean Account of Action

In general terms our topic is to be the sense in which we have control over our actions. A natural starting point is therefore our ordinary conception of human action.

By nearly all accounts, an agent's actions are those of her bodily movements that spring in an appropriate way from her will. Non-bodily movements, and bodily movements that are caused in some other way—the movement of a leaf on a tree that is blown by the wind, say, or the bodily movement of someone who gets tossed about in the surf—are not actions at all. The crucial question that thus arises is what precisely it means to say that a bodily movement 'springs in an appropriate way' from an agent's 'will'. The distinctive feature of philosophical accounts of human action lies in the answer they give to this question.[3]

According to the standard Humean account of human action, for example, an agent's will is to be identified with her *system of desires and means-end beliefs*. The relation in which certain of her desires and means-end beliefs must stand to her bodily movements, for those bodily movements to count as actions of hers, is the *causal* relation. Thus, very roughly speaking, an agent's actions are those of her bodily movements that are caused by her desire for some outcome and her belief that that outcome can be produced by her moving her body in the way she does.

3 David Hume, *A Treatise of Human Nature* (1888). See, eg, Book II, pt III, sec III.

These are the bodily movements over which the agent is supposed to have control.[4]

Unfortunately, however, this rough characterisation is still a little too rough. For suppose that a budding actor desires to sound embarrassed and believes that she can sound embarrassed by saying 'Ugh!', but that this desire and belief cause her to say 'Ugh!' by causing her actually to become embarrassed. She makes an 'Ugh!' sound not as a pretence of embarrassment, but rather as an expression of embarrassment, embarrassment she feels at the prospect of acting on this particular desire and means-end belief. Then it seems that she doesn't have control over what she does despite the fact that her saying 'Ugh!' is caused by a relevant desire and means-end belief.

More precisely, then, the standard Humean account of action tells us that an agent's actions are those of her bodily movements that are caused *in the right kind of way* by her desires and means-end beliefs, where causation in the right kind of way is a matter of the agent's behaviours being not just caused by certain of her desires and beliefs, but also being differentially explainable by them, where differential explanation is a matter of the counterfactual sensitivity of her bodily movements to a whole host of the ever so slightly different desires and ever so slightly different means-end beliefs that she might have had instead. The agent's bodily movement is *counterfactually* sensitive to these slight differences because what *would have happened*, contrary to fact, *would have been different* if, contrary to fact, the agent had had ever so slightly different desires and means-end beliefs.[5]

In the case under discussion, for example, the agent's saying 'Ugh!' counts as an action only if it is not just caused by her desire to sound embarrassed and her belief that she can sound embarrassed by saying 'Ugh!', but it is also the case that, if she had believed that in order to sound embarrassed she would have to say 'Ooooh!', then she would have said 'Ooooh!'; and if she had believed that in order to sound embarrassed she would have to say 'Eeeeh!', then she would have said 'Eeeeh!'; and if she had desired to sound tired as well as embarrassed, and believed that the way to do that was to say 'Ugh!' through a yawn, then she would have said 'Ugh!' through a yawn; and so on and so forth. The reason the budding actor's saying 'Ugh!' as an expression of actual embarrassment doesn't count as an action, and hence isn't something over which she has control,

4 Donald Davidson, 'Actions, Reasons and Causes', reprinted in his *Essays on Actions and Events* (1980) 3–19, is the classic contemporary source of the standard account.

5 Christopher Peacocke, *Holistic Explanation* (1979).

is thus that her saying 'Ugh!' is not differentially explainable by her system of desires and beliefs. Even if she had believed that in order to sound embarrassed she would have to say 'Ooooh!', she would still just have said 'Ugh!'; and even if she had desired to sound tired as well as embarrassed, and believed that the way to do that was to say 'Ugh!' through a yawn, she would still just have said 'Ugh!'; and so on.

More generally, this more precise version of the standard account suggests that the following is a sufficient condition for an agent's having control over what she does: the agent's body moves in a certain way; that bodily movement is caused by a relevant desire and means-end belief the agent possesses; and the agent's bodily movement is also differentially explainable by her desires and means-end beliefs. With this conception of human action and control in the background, let's now ask what it might mean to say that an agent acts on an irresistible impulse, or that she acts but lacks all self-control.

What is an Irresistible Impulse?

A first conjecture would be that an irresistible impulse is an impulse that functions much like embarrassment functions in the situation just described. The reasoning might go like this.

Embarrassment in that case both caused the agent to say 'Ugh!' and would have caused the agent to say 'Ugh!' no matter what small differences we imagine in the desires and beliefs she possesses. In this sense, the impulse does seem both to be irresistible and to cause her to behave in a way that she cannot control. So, by analogy, an impulse that is irresistible must be one which both causes an agent to move her body in a certain way and which would have caused her to move her body in that way no matter what small differences we imagine in the desires she has and the means-end beliefs she has.

The problem with this first conjecture is, however, perhaps already clear. But in order to make the problem with it vivid, let's consider an agent who, depending on how we embellish his story, does plausibly act on an irresistible impulse. Suppose that Bob is a habitual drug user. He desires very strongly to take some heroin in the next short while and believes that, in order to do so, he will have to get some money. He considers the various ways in which he could get the money he needs in the time he has available and concludes that the most efficient method is to break into a certain house and steal it. As a result, let's suppose he breaks into the house and gets the needed money.

Now consider the conjecture. Is it at all plausible to suppose that Bob's desire to take heroin both causes him to move his body in a certain way—that is, in the break-into-the-house-and-steal-a-particular-amount-of-money way—and would have caused him to move his body in that way no matter what small differences we imagine in his wants and means-end beliefs? It most certainly is not plausible to suppose this. For Bob's bodily movement is, after all, an action, from which it follows straight away that it is a bodily movement which is not just caused by his desire for heroin and his belief that the way to get the money he needs is to break into a house and steal it, but is also differentially explainable by his desires and means-end beliefs. In other words, if Bob had thought that the most efficient way to get the money required for heroin was to break into the house a few minutes later than he had originally planned, then he would have broken into the house a few minutes later than he had originally planned; if he had desired to take a slightly higher dose of heroin, then he would have stolen the slightly larger amount of money required for the larger dose; and so on and so forth.

The upshot is thus that the mere fact that Bob acts at all on his desire to take heroin in the next short while and his belief about how that is to be accomplished suffices to falsify the first conjecture. The mere fact that an agent acts at all would seem to guarantee that he exerts a good deal of control over his bodily movement, precisely by ensuring the counterfactual sensitivity of what he does to small changes in his desires and means-end beliefs. Whatever an irresistible impulse is, then, it is nothing much like uncontrollable embarrassment.

A second conjecture therefore suggests itself, and this is that an irresistible impulse is a desire which both causes and differentially explains an agent's action, but which, in addition, is so strong that it is impossible for the agent who possesses it to have had an even stronger desire which would have outweighed it. In the case of Bob, the idea is thus that his desire for heroin is irresistible if it is so strong that it would be impossible for him to have an even stronger desire to do something else instead.

But this conjecture is hard to take seriously. For no matter how strong we imagine Bob's desire for heroin to be, it seems that we can always imagine a desire that is a little stronger. Indeed, it seems that we can often imagine circumstances in which an agent who plausibly has an irresistible impulse has actual desires which, in the right circumstances, would outweigh his impulse. Suppose, for example, that at the moment that he was about to break into the house Bob had seen a swarm of bees flying around inside. Is it supposed to follow from the fact that Bob's desire for

heroin is irresistible that he would have gone ahead and broken into the house anyway? That seems manifestly implausible. But if this is right then not only is it possible for Bob to have a stronger desire than his desire to take heroin, he in fact has such a desire, whether or not his desire is irresistible: the desire not to be repeatedly stung by a swarm of bees.

This brings me to a third and final conjecture, which is that in order to make sense of the idea of an irresistible impulse we will have to go beyond the standard Humean account of action within which we have so far been trying to make sense of the idea. In particular, we will need to recognise the fact that the desires that agents have are themselves often arrived at on the basis of deliberation, that is, on the basis of reflection about what it would be good to do, or what they should do, or what it would be rationally justifiable for them to do. Here, accordingly, we find a further sense in which an agent can exercise control over what she does. To be in control, in this further sense, it suffices that an agent's desires are suitably explainable by her deliberations, that is, by her beliefs about what it would be good to do, or what she should do, or what it would be rationally justifiable for her to do. An irresistible impulse would then be a desire that, in a yet to be specified way, eludes control by these beliefs in the circumstances in which she acts. The further detail that we need to add to Bob's story, in order to establish whether his desire for heroin is or is not irresistible, is thus whether his desire is suitably controlled by his deliberative beliefs.

Going Beyond the Standard Humean Account of Action

At this point, however, we run into a familiar difficulty. For a defining feature of the standard Humean account of action is that beliefs are incapable of playing the kind of explanatory role we have just envisaged for them.

This is not to say that beliefs are inert, on the Humean account. Rather it is to say that, on that account, much as with desires, beliefs are incapable of explaining actions all by themselves. An agent who merely had beliefs is, according to the Humean account, incapable of acting because the mere fact that she believes that the world is a certain way doesn't tell us whether or not she is disposed to make it that way, or some other way. Equally, an agent who merely had desires would be incapable of acting because the mere fact that she is disposed to make the world a certain way doesn't tell us how she thinks the world needs to be changed, or even if it needs to be changed at all, in order to make it that way. To be capable of acting at all,

then, the standard Humean account of action insists that an agent must have both desires and beliefs. But in that case an agent's beliefs about what it would be good to do, or what she should do, or what it would be rationally justifiable for her to do, must be incapable of playing the explanatory role suggested. For, the suggestion goes, such beliefs would have to be capable of both causing and rationalising an agent's having certain desires rather than others all by themselves, and this is something no belief can do.

To respond to this familiar difficulty it seems to me that we must do two things. First, we must say what exactly it is that an agent believes when she believes that she should or shouldn't behave in a certain way, and then, second, we must explain how beliefs with that sort of content are able to play the role we have imagined for them in explaining an agent's desires. What is it about the content of such beliefs that enables them to play that explanatory role? Once we have answered both these questions then it seems to me that we will be in a position to explain the further sense in which an agent can have control over what she does, and this, in turn, will enable us to define the idea of an irresistible impulse.

Let me therefore begin with the first question. What is it that an agent believes when she reflects on her options and comes to the conclusion that it would be good to act in a certain way, or that she should act in that way, or that it would be rationally justifiable for her to act in that way?

I want to approach this question somewhat obliquely by first describing Bob's case in a little more detail. Assume that Bob has two intrinsic desires: a stronger desire that his children fare well and a weaker desire to experience pleasure. Given that he also has various means-end beliefs it follows that, if he were fully instrumentally rational—that is to say, if he were a creature with the ability to perfectly satisfy his intrinsic desires in the light of his means-end beliefs—then he would have extra extrinsic desires as well. That is to say he would have extra desires for things that he doesn't desire intrinsically, but merely as a means to the things that he does desire intrinsically. So let's assume further that, because he intrinsically desires pleasure and believes that taking heroin is pleasurable, Bob would, if he were fully instrumentally rational, extrinsically desire to take heroin, and let's also assume that, because he intrinsically desires that his children fare well and believes that taking heroin will prevent them from doing so (perhaps because he believes that doing so will cause him to neglect them), he would also, if he were fully instrumentally rational, extrinsically desire not to take heroin.

With this background in mind, let's now ask what conclusion Bob might come to if he were to reflect on his options and ask himself what it

would be good to do, or what he should do, or what it would be rationally justifiable for him to do. A natural interpretation of this question now suggests itself, an interpretation according to which we imagine Bob asking himself what he would most want himself to do, in his present circumstances, *if he were fully instrumentally rational*. Moreover, when we interpret the question in this way the answer also becomes clear. For what Bob plainly should do, in this sense—that is, what he would want himself to do, in his present circumstances, if he were fully instrumentally rational—is to refrain from taking heroin. He should refrain from taking heroin because his intrinsic desire that his children fare well is stronger than his intrinsic desire to experience pleasure, and so, if he were fully instrumentally rational, the strengths of his extrinsic desires to take heroin and not to take heroin would simply follow suit.

However even though this is what Bob *should* do, in the sense just explained, it doesn't follow that it is what he *will* do. For what an agent will do is a function of the extrinsic desires she in fact has, not those she would have if she were fully instrumentally rational. Bob's stronger desire that his children fare well will have no impact whatsoever upon his behaviour if it doesn't first combine with a means-end belief to generate an extrinsic desire to do what he believes to be a means to his children's faring well. So even though Bob's desire not to take heroin *should* be stronger than his desire to take heroin—'should' in the sense that it would be stronger if he were fully instrumentally rational—it might not be stronger in fact because he might not be fully instrumentally rational. When Bob deliberates and asks himself what he should do it is therefore a real possibility, a real possibility that we can imagine realised in his particular case, that he will come to the conclusion that he should act in a way in which he has no inclination to act.

Once we see that this interpretation of the claim that an agent should act in a certain way is available, we can readily see that other interpretations are available as well. For, generalising on the basis of this interpretation, the claim that an agent should act in a certain way is plausibly thought to amount to the claim that she would desire that she acts in that way if she had a set of desires that was fully rational *simpliciter*, where being fully rational *simpliciter* is a matter of eluding *all* forms of rational criticism. For example, since we can rationally criticise an agent's desires on the grounds that they are based on inadequate information, so the desires an agent would have if he were fully rational *simpliciter* are those he would have if he were fully informed. And since we can rationally criticise an agent's desires on the grounds that they contribute incoherence to an otherwise coherent desire set, so the desires an agent would have if

he were fully rational *simpliciter* are those he would have if his desire set was maximally coherent. (Indeed, having desires that conform to the principle of instrumental rationality is arguably one dimension along which the coherence of an agent's desire set is measured.) And so we might go on.[6]

An example might help us bring out the way in which these further grounds for rationally criticising the sets of desires that agents have make possible even more glaring cases in which agents deliberate and come to the conclusion that they should act in certain ways, even though they do not desire to act in those ways. Imagine a variation on the case we have been discussing. Bob has just one intrinsic desire: a desire for pleasure. However the reason he has only this one intrinsic desire is complicated. In the past he had another intrinsic desire as well, an intrinsic desire that his children fare well. But at a certain point he fell in with a group of friends who dabbled in drugs for fun and recreation. Though it all began as harmless fun, over a period of time he found his own craving for drugs increased, and as his craving increased he found that he started leaving his children to their own devices more and more so that he could indulge himself. He initially hated himself for neglecting them, but as the neglect increased he managed to decrease the dissonance he felt by telling himself various lies: that he was useless; that his children would be better off without him; that they wouldn't understand his predicament if he told them about it; that they hate him anyway; and so on. The strategy was so successful that in the end he found himself believing the lies that he had told himself. As a result, he simply didn't care as much for his children as he used to. Finally he lost his desire that they fare well altogether.

Given this background, let's now ask what Bob should do in the circumstances he faces, where this is interpreted as his asking himself what he would want himself to do, in his present circumstances, if he were fully rational *simpliciter*, that is, abstracting away from any other requirements that there might happen to be, if he had a desire set that was maximally informed and coherent (where the coherence of a desire set includes conformity to the means-end principle). Taking the story as told at face value it seems to me quite plausible to suppose that the answer to this question is the same as before. On the one hand, even if Bob were fully informed he would still have his intrinsic desire for pleasure, so if he were also fully instrumentally rational then he would also have an extrinsic desire to take heroin as a means to pleasure. But, on the other hand, it is also plausible to suppose that he would regain his stronger desire that his

6 Michael Smith, *The Moral Problem* (1994) ch 5.

children fare well if he were to stop believing all of the lies that he has told himself and immerse himself fully in all of the facts—that he isn't useless; that his children aren't better off without him; that they would understand his predicament if he told them about it; that they still love him; and so on—and, having regained this stronger intrinsic desire that his children fare well, if he were fully instrumentally rational then he would also have a stronger extrinsic desire not to take heroin as a means to preventing his children's being neglected. What Bob would most want himself to do if he had a maximally informed and coherent desire set is thus to refrain from taking heroin. This is what he would most want himself to do if he were fully rational *simpliciter*.

Moreover, note that Bob might well even come to believe this to be so as the result of deliberation. He might, for example, become convinced that he would most want to refrain from taking heroin if he were fully rational *simpliciter* by talking with a trusted counsellor, or a friend. But since being convinced that he would most desire to refrain from taking heroin if he had a maximally informed and coherent desire set is one thing, and knowing what those relevant facts are and being maximally coherent (where this includes being fully instrumentally rational) is quite another, it follows that, notwithstanding his belief, Bob's desire not to take heroin might well not be stronger than his desire to take heroin in fact. When Bob deliberates and asks himself what he should do it is therefore once again a real possibility, one that we can imagine realised in his case, that he will come to the conclusion that he should act in a way in which he has no desire whatsoever to act.

Until now we have been focusing on the first of the two questions distinguished earlier: what is the content of the beliefs an agent forms when, as part of a process of deliberation, she asks herself what it would be good to do, or what she should do, or what it would be rationally justifiable for her to do? The answer we have come up with is that she thereby attempts to form beliefs about what she would want herself to do if she had a set of desires that was fully rational *simpliciter*, where this is a matter of having a set of desires that is maximally informed and coherent (where being coherent includes having desires that conform to the means-end principle). With this conception of the content of the beliefs agents form when they deliberate firmly before our minds, it seems to me that we are now in a position to answer the second question. What is it about the content of these beliefs that enables them to explain the agent's acquisition of corresponding desires?

In order to make matters more concrete, let's ask this question with respect to one of the two scenarios we have been considering. Let's

suppose once again that Bob has just one intrinsic desire, a desire to experience pleasure, but that he used to have a stronger intrinsic desire that his children fare well in the past before he began telling himself all of those lies. In this context let's suppose that Bob deliberates. He consults widely and becomes convinced, after talking with a trusted friend, that he would desire that he stops taking heroin if he had a maximally informed and coherent set of desires. However, to make matters more straightforward, let's suppose that the friend has given Bob no grounds whatsoever for supposing this to be so. He has simply asked Bob to take his word for it, and Bob, trusting his friend as he does, has done just that.

We are thus to imagine that Bob finds himself with the belief that he would desire that he stops taking heroin if he had a maximally informed and coherent set of desires, and the question we must ask ourselves is how this particular belief is supposed to be capable of explaining his acquisition of a desire not to take heroin. What is the mechanism of acquisition supposed to be? The answer I propose is that Bob's belief can explain his acquisition of a desire not to take heroin because considerations of coherence augur in favour of his acquisition of this desire, given that he has the belief he has. The mechanism of acquisition would thus be Bob's quite general non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence.[7]

To see why considerations of coherence look to be the key, consider the following two rather simplified psychologies. One comprises both an agent's belief that she would want herself, in her present circumstances, to act in a certain way if she had a maximally informed and coherent set of desires and, in addition, a desire of hers to act in that way. The other psychology comprises her belief that she would want herself, in her present circumstances, to act in a certain way if she had a maximally informed and coherent set of desires, but does not comprise, in addition, a desire of hers to act in that way. Perhaps it comprises indifference, or aversion to acting in that way. What can we say about these two psychologies, from what we have said about them so far?

The answer seems to me to be plain enough. What we can say is that the first psychology exhibits much more in the way of coherence than the second. For the mere fact that agents fail to have desires, as regards what to do in their present circumstances, that they believe they would have if they had a maximally informed and coherent set of desires, would itself seem to *constitute* a kind of incoherence, or disequilibrium, in their

7 Michael Smith, 'The Coherence Argument: A Reply to Shafer-Landau' *Analysis* (forthcoming).

psychology. The mismatch between such agents' desires about what they are to do in their present circumstances and their beliefs about what they would want themselves to do, in these circumstances, if they had a maximally informed and coherent set of desires bears a striking family resemblance to paradigm cases of incoherence in a psychological state, a family resemblance so striking that we should simply admit that this is a case of incoherence too.

Moreover the fact that this is so is in turn significant. For it is plausible to suppose that rational agents possess a quite general non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence.[8] It is, after all, rational agents' possession of this capacity that explains why they tend to acquire beliefs that conform to the evidence available to them. Moreover it also explains why, when they do not acquire such beliefs, they take themselves to be liable to censure and rebuke. Rational agents quite rightly feel shame when they fail to believe in accordance with the evidence available to them because, given that the evidence dictates that belief, norms of coherence entail that they should have acquired the belief, and because, in the light of the fact that they possess the capacity to acquire the belief, they could have acquired it. They therefore rightly feel shame because they failed to acquire a belief that they should and could have acquired.

Similarly, rational agents' possession of the quite general non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence explains why they tend to acquire desires for the believed means to their desired ends and why, when they do not acquire such desires, they likewise take themselves to be liable to censure and rebuke. Rational agents quite rightly feel shame when they fail to desire the believed means to their desired ends because, given that coherence augurs in favour of the acquisition of such desires, they should have acquired them, and because, in the light of the fact that they possess the capacity to acquire these desires, they could have acquired them. They therefore rightly feel shame because they failed to acquire desires that they should and could have acquired.

It therefore follows that, if I am right that agents who believe that they would desire themselves to act in a certain way, in their present circumstances, if they had a maximally informed and coherent psychology, but then fail to have a corresponding desire, display a lack of coherence in their psychology as well, then the capacity that rational agents possess to

8 Michael Smith, 'A Theory of Freedom and Responsibility' in Garrett Cullity and Berys Gaut (eds), *Ethics and Practical Reason* (1997) 293–319.

acquire and lose psychological states in accordance with norms of coherence has the potential to explain not just why their beliefs tend to evolve in conformity to evidence, and their desires in conformity to their desires for ends and beliefs about means, but also why their desires as regards what they are to do in their present circumstances tend to evolve in conformity to their beliefs about what they would want themselves to do, in their present circumstances, if they had a maximally informed and coherent set of desires.

The question we have been attempting to answer is in what sense we are to suppose that Bob's belief that he would desire that he not take heroin, in his current circumstances, if he had a maximally informed and coherent set of desires, a belief he might form when he deliberates, is capable of explaining his acquisition of a desire not to take heroin in his current circumstances. We now have our answer. Bob's belief is capable of explaining his acquisition of a desire not to take heroin to the extent that Bob is someone who has the non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence. Contrary to the standard Humean dogma, it should therefore be no more puzzling that agents can acquire corresponding desires in the light of their beliefs about what they would desire if they had a maximally informed and coherent set of desires, than it is to suppose that they can acquire beliefs in the light of their appreciation of the evidence for those beliefs, or that they can acquire desires for the believed means to their desired ends in the light of their desires for those ends and their beliefs that the means are means to those ends. The mechanism of acquisition is the same in each case.

The picture we have is thus one according to which agents who are capable of deliberating—that is, capable of having not just desires and means-end beliefs, but of forming beliefs about what they would desire themselves to do if they had a maximally informed and coherent set of desires—and who, in addition, have a quite general non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence, are capable of controlling their behaviour in a further sense. It suffices for an agent to be in control of what she does in this further sense that her body moves in a certain way; that her bodily movement is caused by a relevant desire and means-end belief she possesses; that her bodily movement is counterfactually sensitive to small changes in her desires and means-end beliefs; that her desire is caused by her belief that she would want herself to act in that way in her present circumstances if she had a maximally informed and coherent set of desires; and that her desire is counterfactually sensitive to small changes in her beliefs about what she would want herself to do if she had a maximally informed and

coherent set of desires. (The last condition simply rules out the possibility that the agent desires to do what she believes she would want herself to do in her present circumstances if she had a maximally informed and coherent set of desires as a matter of luck.)

With this further story of control of an agent's desires by his deliberations in the background, we are now in a position to say what it might mean to say that an agent has an irresistible impulse. We are also in a position to clarify the sense in which agents possess the capacity for self-control.

What is an Irresistible Impulse (Again)?

An initial thought might be this. An irresistible impulse is simply any impulse which causes an agent to act, but which isn't caused by her belief about what she would want herself to do in her present circumstances if she had a maximally informed and coherent set of desires; or which, though caused by her beliefs, isn't counterfactually sensitive to small changes in her beliefs about what she would want herself to do if she had a maximally informed and coherent set of desires.

However this can't be quite right. For an agent might well have, and act on, a desire which (say) is not caused by her belief about what she would want herself to do in her present circumstances if she had a maximally informed and coherent set of desires, and yet still have the *capacity* to have and act on the desire that is so caused: 'is not in a state that was so caused' does not imply 'could not have been in a state that was so caused'. Since such a desire would plainly be *resistible*, albeit not *resisted*, it follows that this initial thought does not adequately capture the nature of an irresistible impulse.

More plausibly, then, an irresistible impulse might be characterised as any impulse which causes an agent to act when that impulse is of a kind such that the agent lacks the capacity to have desires of that kind that accord with her beliefs about what she would want herself to do in her present circumstances if she had a maximally informed and coherent set of desires. The idea behind this condition is that an agent might be perfectly capable of desiring in accordance with her beliefs about what she would want herself to do in her present circumstances if she had a maximally informed and coherent set of desires so long as her beliefs are about desires with certain restricted subject matters, while being quite incapable of so desiring when her beliefs are about desires that concern other subject matters: say, drugs, alcohol or gambling.

Note how different this conception of an irresistible impulse is to the conception that was considered, and rejected, earlier. What matters in ascertaining whether an impulse is or is not resistible is not whether the agent could have had some alternative desire that would have outweighed the impulse in question. What matters is rather whether the agent has the capacity, in the circumstances of action she faces, to desire in accordance with her beliefs about what she would want herself to do in these circumstances if she had a maximally informed and coherent set of desires. The mere fact that, if her circumstances were completely different—say, because she had a much stronger competing desire not to be stung repeatedly by a swarm of bees—then she would desire and behave differently is thus neither here nor there.

Are there any irresistible impulses, as just characterised? The question is largely empirical, but it certainly seems to me that we are ordinarily prepared to recognise that there are at least some such impulses as we go about our everyday lives. Addictions are, after all, a common feature of the contemporary world, and what makes a desire into an addiction would seem to be precisely that it meets the condition just characterised: addictions are impervious to deliberative control. This is not, of course, to say that it is easy to prove, either beyond reasonable doubt or according to the balance of probabilities, that some particular impulse is irresistible. But I will leave it to others to determine what evidence might be adduced in support of any particular claim to the effect that some impulse is or is not resistible. My goal here has been more strictly conceptual rather than epistemological.

The Story of Self-Control

It might be thought that the existence of irresistible impulses would provide a rich source of excuses for bad behaviour. In the space that remains, however, I want to voice a note of caution about supposing this to be so. To anticipate, the reason is that, somewhat surprisingly, the mere fact that an impulse is irresistible does not imply that the agent lacks all capacity for self-control.

In deciding whether the existence of an irresistible impulse provides grounds for an excuse, the important point to remember is that there are, for the most part, two quite distinct moments at which agents can exercise such capacity as they have to desire in accordance with their beliefs about what they would want themselves to do if they had a maximally informed and coherent set of desires, or, as this capacity is more colloquially called,

their capacity for self-control. And while it might well be true that agents are often incapable of exercising their capacity for self-control at one of these moments, it is much harder to believe that they so frequently lack the capacity to exercise the capacity at the other moment as well.

In order to see that this is so, note that there are two distinct moments at which we can come to realise that we have the potential to lose control of what we do. Suppose we envisage the possibility, at time t_1, that we will be out of control at time t_2. In other words, suppose we believe, at t_1, that we would want ourselves to act in a certain way at t_2 if we had a maximally informed and coherent set of desires, and believe as well that there is a good chance that at t_2 we lack that desire. The two distinct moments then reflect the possibility that t_1 and t_2 are the *same* time, and the alternative possibility is that they are *different* times.

Let's focus initially on the case in which t_1 and t_2 are the *same* time. To fix ideas, let's consider the particular situation in which Bob has two intrinsic desires, a stronger desire that his children fare well and a weaker desire to experience pleasure, and various means-end beliefs, and on the basis of all these let's suppose that he comes to the conclusion that, if he had a maximally informed and coherent set of desires, then he would have a weaker extrinsic desire to take heroin and a stronger extrinsic desire that he refrain. However, let's also suppose that, faced as he is with availability of heroin right before him, and encouraged as he is by his friends who remind him what great fun he would be missing out on if he refused, Bob becomes instrumentally irrational. His stronger intrinsic desire that his children fare well thus does not transfer its force across the means-end relation, only his weaker desire to experience pleasure does that, and, hence, he finds himself with a stronger extrinsic desire to take heroin rather than refrain.

Now, of course, the mere fact that Bob's extrinsic desire to take heroin is stronger than his extrinsic desire to refrain is no proof that he lacks any *capacity* to have a stronger extrinsic desire to refrain from taking heroin. We therefore need to address the question of his capacity on its own merits. One obvious question to ask, in this regard, is whether any strategy of self-control was available to him. For example, is there something Bob could have thought, or something he could have imagined, such that, if he had thought or imagined that thing then his desire that his children fare well would have transmitted its force across the means-end relation, in which case his extrinsic desire to refrain from taking heroin would have been stronger? Could he, say, have dwelled on the thought that his children depend entirely on him for their well-being, or could he have pictured the disappointment that would appear on their faces if they were watching him take heroin yet again, and, if he had had that thought, or pictured that scene, would this have had the effect of ensuring that his intrinsic desire that his children fare well transmitted its force across the means-end relation?

If we can answer 'yes' to some such question then it seems to me that Bob does, at that time, have the capacity to desire in accordance with his belief about what he would want himself to do if he had a maximally informed and coherent set of desires. If not, then it seems to me that Bob does not have that capacity. In the former case he fails to exercise self-control when the needed exercise of self-control was available to him. In the latter case he fails to exercise self-control, but no such exercise of self-control was available to him in the first place. Accordingly, in the former case we would not suppose that Bob is a candidate for being excused for what he does, whereas in the latter we might well suppose that he is.

In reality, of course, the difference that is being highlighted here will be one of degree. There will be cases in which it is as obvious as can be that there is something that Bob could have thought, or something that he could have imagined, which is such that, if he had thought or imagined that then his desire that his children fare well would have transmitted its force across the means-end relation. In those cases it will no doubt seem too weak to say that it was *merely possible* for Bob to have such a thought, or to engage in such an episode of the imagination, for it will be *astonishing* that he didn't *in fact* have the required thought, or engage in the required episode of the imagination. For example, if in the past Bob has always succeeded in having such thoughts, or in engaging in such episodes of the imagination, then, barring something special about this case, we will think that the possibility of his having the required thought, or engaging in the necessary episode of the imagination, in this case was, as we might say, a *real live possibility*. In other cases, however, though still possible, it will be a far more remote possibility for Bob to have the required thought, or to engage in the required episode of the imagination. For example, if he only very occasionally succeeds in having such thoughts, or engaging in such episodes of the imagination, then we will not be at all surprised that he failed (yet again) on this occasion to have the required thought, or to engage in the required episode of the imagination.

What all of this reflects, of course, is the fact that, in reality, an agent's capacity for self-control *comes in degrees*. What we should really say is thus that an agent, like Bob, might well be excused for doing what he does to the extent that his doing otherwise would have required an exercise of self-control that was beyond him. Even those who have a diminished capacity for self-control are required to exercise such capacity as they

have. But they, too, might be excused when the needed exercise of self-control was beyond their capacity.

But now note that 'might'. So far we have focused on the capacity agents have for *synchronic self-control*. But the mere fact that an agent, like Bob, couldn't have exercised self-control at the moment at which he was vulnerable—the mere fact that he lacked the capacity for synchronic self-control—does nothing to show that he couldn't have exercised self-control at *some other time*. Let's therefore consider Bob's situation once again, but this time let's pull back to an earlier time at which Bob is at home with his children, fixing their dinner, suffering from no instrumental irrationality. Indeed, let's suppose that he has recently thought through his situation and has resolved to give up taking heroin, because he has foreseen the terrible effect that his heroin use will have upon his children's lives. At that very instant, however, his friends call him on the telephone and invite him over. They tell him that they have just purchased some heroin and that they would like him to join them for an evening of fun and recreation. Suppose they say that even if he doesn't want to take heroin, he should come over anyway just to have a few beers and a chat. What should Bob do?

The crucial point to note is that, at this moment, Bob has available both of the following beliefs. First, he has available the belief that, if he had a maximally informed and coherent set of desires then he would want himself, at the later time, to refrain from taking drugs. Second, he also has available the belief that, if he were to go along to his friends' house, resolving only to have a few beers and a chat, the prospect of taking heroin would make him instrumentally irrational and he would end up taking heroin despite his resolve: this, after all, is what has happened to him time and time again.

We are therefore quite within our rights to suppose that, at the earlier time, Bob can well envisage the prospect of his losing control of himself at the later time, and so, given that he desires more strongly that his children fare well, at that earlier time, and only less strongly that he experiences pleasure, it follows that, if Bob had envisaged that prospect, then, given that he is instrumentally rational, he would have extrinsically desired not to join his friends that evening. Moreover, in acting on this desire he would have been exercising a distinct kind of self-control: *diachronic self-control*. Bob possesses the capacity to exercise diachronic self-control at the earlier time because, being able to foresee that he would lose control in the future if he allowed certain circumstances to obtain, he is able to so construct his future circumstances as to ensure that those circumstances do not obtain.

The upshot is thus that, even if Bob is unable to exercise synchronic self-control—that is, even if he is not able to think or picture anything that would ensure that he doesn't suffer from instrumental irrationality and the desire to take heroin, when faced with the prospect of actually doing so—he might still not be excusable. He might not be excusable because, to be excusable, it would have to be the case that there was no prior moment at which Bob could have exercised diachronic self-control.

Conclusion

I said at the outset that my aim in this paper was to examine two ideas crucial to a proper interpretation of the McNaghten Rules: the idea of an irresistible impulse and the correlative idea of an agent's lacking self-control. The main findings can now be summed up as follows.

Though possession of desires and means-end beliefs that cause and differentially explain an agent's bodily movements suffices for agents to be in control of what they do in one sense, it does not suffice for their being in control in another, and more important, sense. For agents are in control in this more important sense when their desires are suitably responsive to their deliberations, that is, to their reflectively formed beliefs about what they would want themselves to do if they had a maximally informed and coherent set of desires. The capacity for self-control is one aspect of this responsiveness. It is the capacity rational agents possess to have desires corresponding to those they believe they would have if they had a maximally informed and coherent set of desires, a capacity which, in turn, is an instance of a more general capacity they have to acquire and lose psychological states in accordance with norms of coherence.

Armed with this definition of the capacity for self-control we can define the idea of an irresistible impulse. An irresistible impulse is an impulse which eludes an agent's exercise of his capacity for self-control. The distinction between synchronic and diachronic exercises of self-control is, however, crucial at this point. For to be completely irresistible an impulse must be more than one which an agent is unable to conquer synchronically: in other words, more must be true than that no feat of the imagination or thought which was within the agent's reach at the time could have stopped the impulse from having its effect at the moment at which he suffers it. An agent who could have foreseen that he would be out of control if he were to find himself in certain circumstances in the future, and who failed to take such steps as were available to him to ensure that those circumstances did not arise, though he may well suffer from a

synchronically irresistible impulse, does not suffer from an impulse which is diachronically irresistible. Such an agent is not excusable for doing what he does, notwithstanding the fact that he acts on a desire that is, in a sense, irresistible.

4 Intention and Agency

GRANT GILLETT

In what way does the nature of intention reveal more than a physical description of bodily movements and engage our thought with the character of the agent who acts upon that intention? I will argue that to answer this question we have to achieve some clarity on four holistically related concepts. First we have to speak of the *person as an integrated rule follower* to understand the way the agent forms mental content. Second we have to speak of *mental content* to understand *the identity of an action* (a third key concept) and fourth we have to consider the individual who composes *a lived narrative* which is more or less coherent to understand the agentic origins of intentional action.

What is an Intention?

There is clearly a difference between that casual Mediterranean shrug of the shoulders and out-turning of the hands that *betrays* or inadvertently reveals the fact that one's body has been inscribed by a particular discursive context and the self-conscious production of that same gesture for effect. The difference is one of intent. If we were tempted by a certain philosophical view, we might say that in the latter case the gesture was caused by an explicit intention in service of a motive or project of conveying to one's audience that one's character has been infected by a kind of Mediterranean ambience or colouring to the soul. But this view seems too deliberative to do justice to the more spontaneous version in that such a reading would threaten the authenticity of the performance it is trying to explain. So this leads us to another question.

What is the relation between a particular intention and the act which it informs? Since John Stuart Mill, the relation has been conceived to be one