Three Methods of Ethics (1997, with M. Baron and M. Slote) and *The Common Mind: An Essay on Psychology, Society and Politics* (1993). He also has a new book in press *A Theory of Freedom: From Psychology to Politics* (2001).

Arthur Ripstein is Professor of Law and Philosophy at the University of Toronto. He was Laurance Rockefeller Visiting Fellow at Princeton in 1995-96, and was a Connaught Fellow in 2000. He is the author of *Equality, Responsibility and the Law* (1999) and co-editor of *Law and Morality* (1996, 2nd edn, 2001), and *Practical Reason and Preference* (2001). He has also published numerous articles in legal and political philosophy. He is an associate editor of the journal *Ethics*.

Michael Smith is Professor of Philosophy at the Research School of Social Sciences, Australian National University. Author of *The Moral Problem* (1994) and editor of *Meta-Ethics* (1995). His main research interests include moral and political philosophy, philosophy of law, moral psychology, and philosophy of mind and action.

Jane Stapleton is Professor within the Law Program of the Research School of Social Sciences, Australian National University, before which she was Fellow in Law at Balliol College, Oxford. Currently she is Adviser to the *Restatement (Third): General Principles* project and is writing about causation and comparative product liability.

1

RESPONSIBILITY AND SELF-CONTROL

*Michael Smith**

In "Being Responsible and Being a Victim of Circumstance", Tony Honoré tells us:

> "Before imposing sanctions or attaching blame, law and morality requires something more than that the person concerned is responsible for what they have done. One further requirement, common to both . . . is that in the circumstances the agent had the capacity to reach a rational decision about what to do. When this capacity is present, blame for bad behaviour is appropriate and criminal liability may, depending on the state of the law, be imposed. But, though capacity has often to be treated as an all-or-nothing matter, since an offender must be found guilty or not guilty, in real life our ability to decide rationally is a matter of degree. So different degrees of blame, punishment and censure correspond to the extent to which the agent's capacity is impaired." (Honoré, 1998: 138)

Though Honoré is doubtless right about this, the fact that we impose sanctions and attach blame in this way raises several difficult questions. My aim in this paper is to raise some of these difficult questions, and then hopefully to suggest some answers.

The general idea, let's agree, is that the only people we see fit to sanction and blame are those who have rational control over their conduct. We no more sanction and blame those who have no rational control over their conduct—those who are, in Honoré's phrase, "victims of circumstance"—than we sanction or blame floods or earthquakes or lightning bolts for the harm for which they are responsible. The main question on which to focus, then, is what precisely it means for someone to have rational control over their

* An earlier version of this paper was read at *Responsibility in Law and Ethics*, a symposium held in honour of Tony Honoré at the Research School of Social Sciences, Australian National University, November 1999. I would like to thank all those who participated in the discussion at this symposium, but especially Tony Honoré. The paper builds upon ideas that first appeared in "Frog and Toad Lose Control" (Kennett and Smith, 1996). Ss. 2, 3 and 4 include a translation of material that first appeared in "Quelques énigmes concernant le contrôle de soi" (Smith, forthcoming). I am grateful to the editor of *Philosophiques* for his permission to include that material here. Finally, I would like to thank Lloyd Humberstone for a crucial suggestion, and Peter Cane and John Gardner for their very helpful comments on the final draft.

conduct, and, in particular, how it can be that this ability, and the consequent sanctions and blame, comes in degrees.

So as not to muddy the waters with too much unnecessary controversy, my discussion will focus on a particular case in which issues of rational control arise, but which raises no issues of legal or moral significance. My aim is, if you like, to examine and describe the psychological structures required for an agent to possess and exercise the capacity for rational control in a non-controversial case, and then simply to assume that those same psychological structures will be in place, playing much the same roles, in the more controversial cases. With that in mind, here is the case on which I will focus.

Each day on his way to work, John stops at the local supermarket to buy his day's supply of chocolate bars. Despite his love of chocolate, he knows that he shouldn't eat as much chocolate as he does. Given that he is over forty and does no exercise, he realises that the amount of chocolate he consumes simply adds to an already significant weight problem. But his belief that eating so much chocolate will make him fat and induce heart disease simply does not move him. Try as he might, John cannot control himself—or so he says. He buys and eats chocolate notwithstanding his beliefs about what he should do. With this case before us, let's turn to consider the various questions that arise.

1. How can those who intend to act in a certain way be said to lack self-control when they subsequently act in that way?

Note a crucial feature of John's case, as described. John says that when he eats chocolates he is out of control. But it would surely be quite incredible to suppose that he does not intend to eat chocolates. Indeed, as described, it seems that John quite evidently has a standing intention to eat chocolate each day. This provides us with a first puzzle. The puzzle arises because the concept of intention would seem to bring a concept of rational control in its wake.

The constitutive role of an intention is, after all, to ensure that agents act over time in accordance with a plan (Bratman, 1987). Thus, for example, the constitutive role of John's intention to eat chocolate each day is to ensure that John's behaviour, over time, fits in with an overall plan of action that sees him eating chocolate each day. His intention to eat chocolate is thus what ensures that he leaves for work in time to stop at the supermarket before he begins his day's work; it is what ensures that he takes the route to work that goes via the supermarket that sells chocolate; it is what ensures that he has some money with him when he walks into the supermarket; it is what ensures that he takes

the chocolate he buys with him to his office, and doesn't leave it in the car; and so on and so forth.

The mere fact that John has an intention to eat chocolate each day therefore guarantees that he exercises quite a lot of control over his behaviour by placing quite severe constraints on the way in which he conducts himself over time. Very roughly, it ensures that, whatever else he does over time, he does not act in ways inconsistent with his eating chocolate. Moreover, it ensures that he does the things that he needs to do in advance for his subsequent eating of chocolate. But now the puzzle should be evident. For how can we suppose that John lacks rational control when he plainly exercises so much rational control?

The solution to this first puzzle lies in recognition of the fact that, though there is a sense in which John exercises control over his behaviour simply in virtue of having a standing intention to eat chocolate, there is also a sense in which he doesn't. The reason is that intentions themselves are appropriately arrived at as a result of deliberation. When they deliberate, agents reflect on what it would be good to do, or what they should do, or what it would be rationally justifiable for them to do, and, on the basis of these reflections, they form their intentions. But what cases like John's bring out is that an agent's intentions do not invariably answer to the considerations that he takes into account when he deliberates. Sometimes, as in John's case, an agent may intend to act in one way even though, when he reflects, he commits himself to the view that it would be bad to act in that way, or that he should not act in that way, or that his doing so would not be rationally justifiable.

Here, then, we find a residual sense in which an agent may fail to control his behaviour despite the fact that he intends so to behave. For an agent to be in control of himself when he behaves, it is not enough that his behaviour conforms to his intentions. His intentions must in turn conform themselves to his deliberations, that is, to his beliefs about what it would be good to do, or what he should do, or what it would be rationally justifiable for him to do. John thus looks to be a likely candidate to lack control, in this sense—control of himself—because his behaviour, though controlled by his intentions, is not controlled by intentions that are in turn controlled by his deliberations.

2. How can we intentionally act in ways we believe we should not?

The solution to the first puzzle about rational self-control makes an important assumption. It assumes that agents can have beliefs about what it would be good to do, or what they should do, or what it would be rationally justifiable

to do. But what exactly are these beliefs about? What would make them true or false? Moreover, how are such beliefs supposed to impact on our actions?

Thomas Hobbes provides the orthodox answer to this question.

> ". . . whatsoever is the object of any mans Appetite or Desire; that is it, which he for his part calleth *Good*: And the object of his Hate, and Aversion, *Evill*; And of his Contempt, *Vile*, and *Inconsiderable*. For these words of Good, Evill, and Contemptible, are ever used with relation to the person that useth them: There being nothing simply and absolutely so; nor any common Rule of Good and Evill, to be taken from the nature of the objects themselves . . ." (Hobbes, 1651, part 1, ch. 6: 120)

Hobbes's idea is that our beliefs about what is good and bad can impact on our desires and aversions—which, in turn, impact on our intentions—because these beliefs are beliefs *about* our desires and aversions. However, if Hobbes is right about the meaning of "good" and "bad", then it turns out that the distinction made in solving the first puzzle about self-control is a distinction without a difference.

John says that he shouldn't eat so much chocolate. But, translating what John says into Hobbes's terms, what he says is, apparently, that he is averse to eating so much chocolate, or desires not to. But, as the story has been told, not only is this manifestly false, John knows it to be manifestly false. John is not at all averse to eating so much chocolate. Rather, he has very strong desire to eat chocolate, a desire so strong that he has a standing intention to do so, and in recognition of which he reaffirms his view that he shouldn't be eating so much. The Hobbesian translation of John's claim that he shouldn't eat so much chocolate thus casts doubt on the possibility of distinguishing claims about what an agent desires to do or intends to do from claims about what it would be good for him to do, or what he should do, or what it would be rationally justifiable for him to do.

Here, then, is the second puzzle about self-control. We have to give an account of what the "should" means when, in cases in which we lack control, we say that we know we desire to act in ways that we shouldn't. But the orthodox account, Hobbes's account, will not do. It suggests, falsely, that we contradict ourselves in saying what we say.

But nor, importantly, will the obvious variations on Hobbes's account do either. For example, it will not do to suggest that when John says that he shouldn't eat so much chocolate, what he is saying is that he has a desire that his first-order desire to eat chocolate not be effective: that he would prefer that a first-order desire to refrain from eating chocolate be effective in action instead (Frankfurt, 1971). For, as Gary Watson has pointed out, an agent's

second-order desires are simply further desires he possesses (Watson, 1975). They have no special status that would allow them to give content to claims about what we *should* do.

Thus, though it is certainly true that if John has a first-order desire to eat chocolate and a second-order desire that a first-order desire to refrain from eating chocolate be effective in action instead, then the desires that he has will not be cosatisfiable, there would seem to be no reason why it is the first-order desire John has to eat chocolate that should be changed to match his second-order desire, rather than his second-order desire that should be changed to match his first-order desire. In other words, there seems to be no reason why John's lack of control should reside in his possession of a conflicting first-order desire, as opposed to residing in his possession of a conflicting second-order desire. The solution to the second puzzle about self-control thus requires a more radical departure from Hobbes's view.

Let's assume, for the moment, that our desires and aversions can sometimes be the product of irrationality. On that assumption it is simply implausible to suppose, as Hobbes does, that to say that something is good or bad, or that it should or shouldn't be done, is to say that we desire or are averse to it. It is implausible because we would all readily agree that at least some of our desires and aversions are desires and aversions that we should not have in a totally uncontroversial sense, that is, in the sense that we would not have them if we were fully rational (Smith, 1994; 1995).

What would still seem plausible, however, is a modified form of Hobbes's view, a modification in the spirit of the Enlightenment idea, due to Kant, that our desires and aversions must themselves be formed via rational processes. On this view, what each of us, for our own part, calls "good" and "bad" is still a matter of a relation that we stand in to the things that we call "good" and "bad", but the relation is not that of being something that we *actually* desire, or to which we are *actually* averse, but is rather that of being something that we *would* desire, or to which we *would be* averse, *if we were in a more fully rational state*: that is, if we had desires and aversions that eluded all forms of rational criticism.

The question we must ask is thus whether the assumption that our desires and aversions can be liable to rational criticism is correct. And the answer must surely be that it is. Imagine, for example, that John has both a desire for pleasure and a stronger desire for health, and that he believes both that eating chocolate will lead to pleasure and that refraining will lead to health, but that only the weaker desire for pleasure transmits its force across the means–end relation. In that case John's resultant desire to eat chocolate would be liable to rational criticism on the grounds that it is the product of

instrumental irrationality. John might then be able to say quite truly that, notwithstanding his desire to eat chocolate, it would be more desirable for him to refrain from doing so: that is, that he would more strongly desire himself to refrain if he were fully rational.

Or imagine instead that John has both a desire for pleasure and a weaker desire for health, and that he believes both that eating chocolate will lead to pleasure and that refraining will lead to health, but that the relative strengths of his desires would change if his desire set as a whole was more coherent and unified. In that case John's resultant desire to eat chocolate would be liable to rational criticism on the grounds that it is the product of the incoherence or disunity of his desire set as a whole. John might then once again be able to say quite truly that, notwithstanding his desire to eat chocolate, it would be more desirable for him to refrain from doing so: that is, that this is what he would more strongly desire himself to refrain if he were fully rational.

Or imagine instead that though John has only a desire for pleasure and a belief that eating chocolate will lead to pleasure, and though his desires are not a product of either means–end irrationality, or incoherence or disunity in his desire set as a whole, that he is ignorant of certain facts, and that, if he weren't ignorant of those facts, he would have a much stronger desire to be healthy. Imagine further that this desire would be so strong that, given that he believes that refraining from eating chocolate will lead to health, he would then desire even more strongly not to eat chocolate. In that case John's resultant desire to eat chocolate would be liable to rational criticism on the grounds that it is the product of ignorance. John might then, again, be able to say quite truly that, notwithstanding his desire to eat chocolate, it would be more desirable for him to refrain from doing so: that is, that this is what he would more strongly desire himself to do if he were fully rational.

Here, then, we find a natural interpretation of what John is saying when he says that he should stop eating so much chocolate, notwithstanding the fact that his strongest desire is to do just that. He is expressing his belief that this is what he should do all things considered, and this, in turn, is a belief about what he would *most* want himself to do if he were fully rational: that is, if he had knowledge of all the relevant facts, was fully instrumentally rational, and if his desire set as a whole was maximally coherent and unified. This is a natural interpretation of what John says because we are all familiar with situations in which we have no desire at all to act in the way we believe we would act if we were fully rational. Failures of memory and imagination, ignorance, means–end irrationality, incoherence, and all manner of other non-cognitive personality disorders as well, can readily cause us to lose desires that we would have in a more fully rational state, or cause us to have desires that we would

not have in a more fully rational state. These sorts of failures of reason are thus what explain our need to exercise rational self-control.

3. How can an agent both need to exercise self-control and succeed in doing so?

The second puzzle was to explain what the "should" means when we say, as John says, that though we may desire to act in certain ways in certain circumstances, when we think that we need to exercise self-control in those circumstances what we also believe is that we should not act in those ways. The third puzzle is to make some coherent sense of the idea that the needed exercise of self-control is so much as possible.

The source of the problem here lies in a truism in the philosophy of action, a truism popularised by Donald Davidson (1970). According to the truism, if an agent desires most to act in a certain way, and he believes himself free to perform that action, then if he tries to perform any action at all, that is the action he will try to perform. Thus, according to the truism, if right here and now I desire most to continue writing, and I believe myself able to do so, then that is what I will try to do, if I try to do anything. Of course, I might not try to do anything. I might faint, or fall into a coma, or drop dead. But if I do try to do anything, then what I will try to do is to continue writing.

Though the truism no doubt requires more careful formulation than I have given it here, I hope that it does sound truistic. It certainly should, given the ways in which the concepts of desire and belief and action are interdefined. An action is, after all, simply defined as a doing that is the causal upshot of a desire and belief pair, and in cases in which an agent has a variety of desires and beliefs, what she wants most to do is simply defined to be the action that is the object of the desire and belief pair that has the greatest causal power. Despite the fact that the truism requires more careful formulation, then, it seems safe to suppose that it, or something along similar lines, is indeed a truth.

Given the truism, however, it is hard to see how anyone who needs to exercise self-control could ever succeed in doing so. For, as we have seen, an agent who needs to exercise self-control desires most to act in one way while believing that he should act in another. But if an agent desires most to act in one way while believing that he should act in another, the truism tells us that if he tries to do anything at all, then what he will try to do is to act in the way he most wants to act, not in the way he believes he should. Indeed, since the truism purports to be a conceptual truth, it seems to follow that it

is logically impossible for someone who needs to exercise self-control to succeed in doing so. For that would require him both to desire most to act in one way, and yet to desire more strongly to act in the way required for self-control instead, and that is an out and out contradiction in our description of the agent in question.

Return to the case of John, just to drive the point home. John desires most to eat chocolate, but believes he shouldn't. According to the truism, however, it follows that if John tries to do anything at all then he will try to eat chocolate. But if John is going to try to eat chocolate, if he tries to do anything at all, then he evidently isn't going to try to exercise self-control. Though we have succeeded in explaining why John needs to exercise self-control, it thus seems that the needed exercise is a logical impossibility. This is the third puzzle about rational self-control. How are we to solve it?

It seems to me that this puzzle arises because of assumptions that need to be made explicit and then questioned. Note, for example, that there are at least two quite distinct times at which we can exercise self-control. Suppose we envisage, at time t_1, that we will be out of control at time t_2, at least absent an exercise of self-control. The two distinct times at which we can exercise self-control reflect the fact that t_1 and t_2 might be the *same* time, or *different* times. The puzzle, on the other hand, arises only on the assumption that t_1 and t_2 are the same time.

Imagine that t_1 is an earlier time and that t_2 is a later time. We are then in a position to ask ourselves, at t_1, what we most want to do at t_1, and the answer might well be that, believing as we do that at the later time t_2 we have the potential to lose control, what we most want to do is to ensure that at t_2 we do not lose control. And here, accordingly, is one completely straightforward way in which we can perform an action, and so exercise control over our own subsequent actions, at least provided we are not out of control at t_1. We can exercise self-control diachronically, at the earlier time, by so arranging the circumstances of action that we will face at the later time so as to remove the possibility of our then losing control.

Thus, for example, if at an earlier time when, say, he was instrumentally rational, John had foreseen that he would no longer be instrumentally rational when passing the supermarket on the way to work—perhaps the sight of the supermarket makes especially salient the possibility of buying chocolate, and this causes him to become instrumentally irrational and only then to intend to eat chocolate—then, if he had most wanted to do so, he could have made sure that he was unable to act in an instrumentally irrational way at that later time by, say, ensuring that the only option available to him then would be the one that he would have desired if he had been fully instrumentally

rational. He could have driven to work by a different route, say, so that he didn't ever get to see the supermarket; or he could have so arranged things that he had no money in the car on the way to work; or he could have taken someone with him in the car who would talk him out of stopping to buy chocolate; and so we could go on. In this way he could have ensured at the earlier time that, despite his potential to act in an instrumentally irrational way at the later time when driving past the supermarket, that potential was never realised.

In short, then, one solution to this third puzzle about self-control lies in the mundane observation that diachronic self-control is possible. When we exercise diachronic self-control our trying not to do what we most want to do may indeed require that our strongest desire is not to act on our strongest desire. But there is no contradiction involved once we see that these desires are had at different times. When we exercise diachronic self-control at t_1 our strongest desire at t_1 may be that we cause ourselves not to act on what will be our strongest desire at the later time t_2. The trick lies in the fact that the exercise of diachronic self-control at t_1 itself ensures that we are unable to act on our strongest desire at t_2 by making it the case that the only acts available at that later time are acts that satisfy weaker desires.

4. How can an agent both need to exercise self-control and succeed in doing so at the very moment of vulnerability?

What if t_1 and t_2 are the same time? At the very moment that he desires most strongly to eat chocolate, is John able not to act on this, his strongest, desire? This is the fourth puzzle about self-control. Granting that diachronic self-control is possible, is synchronic self-control possible too? Is it possible for an agent both to need to exercise self-control and to succeed in doing so at the very moment of vulnerability?

Some people are sceptical about the possibility of synchronic self-control. One reason for their scepticism is that many examples of what appear to be cases of synchronic self-control turn out, on closer inspection, to be cases of diachronic self-control where the times in question are just very, very close together. Thus, for example, imagine a case in which if John catches a glimpse of the supermarket where he regularly buys all of his favourite chocolates on his way to work then that glimpse will cause an irrational shift in his desiderative profile. It will cause him to become instrumentally irrational, say, and so to desire most strongly to stop and buy some chocolate, whereas before he had the glimpse he had no such desire. If that is right then John might well do something to prevent himself from ever having that glimpse just a moment before he

has it. He might look the other way, or shield his eyes with his hands, or distract himself, or whatever. But though, in such a case, it would be natural to describe John as having pulled himself together at the very moment of vulnerability, the fact is that we imagine John's pulling himself together an instant before he would otherwise have lost control. It is therefore a case of diachronic, not synchronic self-control, and, as such provides us with no embarrassing contradiction in our description of what John does. We can simply imagine John wanting most to prevent himself from having a glimpse of the supermarket, and acting on this desire, an instant before his desires would otherwise have changed.

What cannot be imagined, however, is a case in which an agent, at one and the same time, both wants most to act in one way—the way required for a loss of control—and yet wants even more to act in another way—the way required for an exercise of self-control. John, for example, cannot at one and the same time both want most to eat chocolate and want even more to prevent himself from eating chocolate, for that is an out and out contradiction in our description of John. Sceptics about synchronic self-control quite rightly emphasise this point. The question we must ask, however, is whether the exercise of properly synchronic self-control requires any such thing.

An example might help. Suppose, as seems common enough, that John exercises self-control by having certain thoughts or engaging in certain imaginings. At the very moment at which he wants most to eat a chocolate, suppose he thinks of the chocolate that he is about to eat as a lump of fat, and that he imagines that lump of fat curdling in his stomach after he eats it. This imaginative exercise might well prevent John's desire to eat a chocolate—which by stipulation is the strongest desire he has at that instant—from having its characteristic effect. The exercise of self-control counts as synchronic, because it happens at the very time at which John desires most to eat a chocolate. The question we need answered, however, is what the cause of those imaginings is supposed to be.

The strongest desire John had at the very moment at which he engaged in those imaginings didn't cause it, that's for sure. For, by stipulation, what he wants most to do at that very moment is to eat a chocolate, and his imaginings simply undermine that. Nor did the strongest desire he had an instant earlier cause it either. For we can plainly imagine that what he most wanted to do an instant earlier was also just to eat a chocolate. But in that case, what did cause those imaginings? The answer seems to me to be implicit in what we said at the very beginning about what the "should" means when John says that he knows he shouldn't eat so much chocolate.

Remember, John needs to exercise self-control because, though he believes that he shouldn't eat chocolate, he wants most to do so. As I have already

argued, this belief of John's—his belief that he shouldn't eat chocolate—is itself in turn simply the belief that he would most want himself not to eat chocolate, in his present circumstances, if he was fully rational: that is, if he was fully informed; if he had the set of desires he would have if he were fully instrumentally rational; if his desires as a whole formed a maximally coherent and unified set; and so on and so forth. John's problem, then, is that the strongest desire he actually has fails to accord with the strongest desire he believes he would have in this more fully rational and coherent state of mind.

But now consider the following two psychologies. One comprises both an agent's belief that he would want himself, in his present circumstances, to act in a certain way if he had a maximally informed and coherent set of desires and, in addition, a desire of his to act in that way. The other psychology comprises his belief that he would want himself, in his present circumstances, to act in a certain way if he had a maximally informed and coherent set of desires but does not comprise, in addition, a desire of his 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 just what we have said about them so far?

The answer seems plain enough. What we can say is that the first psychology exhibits more in the way of coherence than the second. 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, itself constitutes a kind of incoherence, or disequilibrium, in their psychology. It constitutes a kind of incoherence or disequilibrium because these agents fail by their own lights. Note that this is not to say that the desire that they believe they would have if they had a maximally informed and coherent set of desires would indeed be an element in such a set. The point is rather that one source of incoherence in the psychology agents have lies in the mismatch between their 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.

The fact that this is so is in turn very significant. For it is independently plausible to suppose that rational agents possess a quite general nondesiderative capacity to acquire and lose psychological states in accordance with norms of coherence (Pettit and Smith, 1996). It is rational agents' possession of this capacity that explains why, for example, 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.

If I am right, however, that agents who believe that they would desire themselves to act in a certain way, in their present circumstances, if they had a maximally coherent psychology, but then fail to have a corresponding desire, display a lack of coherence in their psychology as well, then it seems to follow that the capacity rational agents possess to acquire 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.

In agents who never fail to exercise this capacity we might well expect to find that 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 cause, or at any rate causally sustain, their having of corresponding desires quite generally. They never need to exercise self-control because they never find themselves believing that they would desire themselves to act in one way, in their present circumstances, if they had a maximally informed and coherent set of desires, while yet failing to desire to act in that way. But in agents who aren't as superhumanly coherent as that, we might well expect to find that they possess back-up capacities, capacities that enable them to get back on track when their desires fail to match their beliefs about what they would want themselves to do, in their present circumstances, if they had a maximally informed and coherent desire set. We might expect them to be

disposed, for example, to engage in certain processes of thought or imagination which prevent their divergent desires from having their characteristic effect, and which cause them to have desires that lead them to do what they would have wanted themselves to do if they had had a maximally informed and coherent set of desires instead.

Here, then, lies the explanation of how, despite the fact that he wants nothing more than to eat chocolate, John can none the less imagine chocolate to be a mere lump of fat curdling in his stomach. He can engage in that imaginative exercise because he possesses a quite general non-desiderative capacity to acquire and lose psychological states in accordance with norms of coherence, and that capacity, when exercised, amounts to no more or less than his thinking such thoughts and engaging in such imaginings as will restore his psychology to a more coherent state given its present state of incoherence. In short, John imagines chocolate to be a lump of fat curdling in his stomach because he is a rational creature and that is a rational thing for him to imagine at that time. It is a rational thing for him to imagine at that time because it causes him to lose his desire to eat chocolate—the desire whose presence makes for incoherence in his psychology—and instead causes him to desire most to do what he believes he would most want himself to do if he had a maximally informed and coherent and unified set of desires. In this way John restores coherence to his psychology.

The fourth puzzle about self-control arises because of the apparent logical impossibility of properly synchronic exercises of self-control: that is, cases in which agents both need to exercise self-control and succeed in doing so at the very moment of vulnerability. These cases seemed to be impossible because their description seemed to involve a contradiction. It looked like agents had to both want most to act in the out-of-control way, and yet to want even more to act in the way required for the exercise self-control. The solution to this puzzle lies in the fact that exercises of properly synchronic self-control aren't caused by desires and means–end beliefs. When agents exercise properly synchronic self-control they engage in various thought processes and imaginings that are caused by a non-desiderative capacity they possess, in particular, by the capacity to acquire psychological states in accordance with norms of coherence.

5. What makes it true that we can exercise rational self-control, when we can?

I have suggested that despite the fact that John wants nothing more than to eat chocolate, he can still exercise self-control. He has the capacity to do so,

even though he might fail to exercise that capacity. But what exactly makes it true that John can exercise self-control when he fails to exercise it? What is the difference between an unexercised capacity for self-control and no capacity for self-control at all? This is the fifth puzzle about self-control.

In "Can and Can't" Tony Honoré offers the makings of an answer. Honoré suggests that we need to distinguish between two "cans": "can" (particular) and "can" (general). In these terms, to say of John that he can (particular) exercise self-control is, according to Honoré, to say of him that he does or will exercise self-control.

> "[S]uccess or failure, on the assumption that an effort has been or will be made, is the factor that governs the use of the notion. If the agent tried and failed, he could not do the action: if he tried and succeeded, he was able to do it. If he will fail however hard he tries, he cannot do it; if he will succeed provided he tries, he can". (Honoré, 1964: 144)

What makes it true of John that he can (particular) exercise self-control is thus that he succeeds in doing so. His failure to exercise self-control suffices for the truth of the claim that he can't (particular). Plainly, this can't be what makes it true of John that he can exercise self-control when he fails to exercise that capacity.

This is where Honoré thinks that we need to appeal to the idea of what an agent can (general) do. He suggests that "can" (general) is most commonly used in connection with types of performance in order to claim a general competence, or ability, or skill.

> "[A] condition sufficient for asserting 'he can (general) do such-and-such a type of action' is that, when the agent tries, he normally succeeds in doing an action of that type. . . . 'Can' (general), when used of particular actions, differs from 'can' (particular), when used of particular actions, in that its correct use does not depend on actual or prospective success or failure." (Ibid.: 145–6)

Thus, according to Honoré, the sense of "can" in which we are interested—the sense in which John can exercise self-control, though he might fail to do so—is presumably the sense of "can" (general). What makes it true of John that he can (general) exercise self-control is thus that he usually does exercise self-control when he tries to do so.

Though, as we will see, I have some sympathy with Honoré's idea, it seems to me that we must reject his suggestion as it stands. True enough, our best *evidence* for the truth of the claim that John can exercise self-control, when he fails to do so, is very often that he usually succeeds when he tries. But it is hard to believe that the truth of the claim that John can exercise self-control, when he fails to do so, *consists* in the fact that he usually does when he tries. It is hard

to believe for the simple reason that John's possession of the ability to exercise self-control doesn't seem to require that he exists for longer than a moment.

In order to see that this is so, consider the following thought experiment. Imagine that we have the technology to create an exact replica of John in a laboratory, a molecule for molecule duplicate. Suppose that the replication process takes place, but a few moments after the process is complete, through some mishap or other, the replica is destroyed. To my mind, so long as we are convinced that the creature we created really was a molecule for molecule replica of John, we wouldn't hesitate for a moment in ascribing to him all of the psychological states and capacities that we would be willing to ascribe to John himself, including John's capacity for self-control.

Thus, suppose that during the moments the replica existed we offered both John and his replica a piece of chocolate and that they both ate it. Let's suppose further that we are willing to say of John, at that time, that he could have exercised self-control by refraining from eating the chocolate and that, what's more, this is true. It seems to me that we should then be willing to say exactly the same thing of John's replica, and that what we say of John's replica should in that case be true too. But since John's replica doesn't exist long enough for it to be true of him that he normally succeeds in exercising self-control on such occasions when he tries, it follows that what makes it true to say of him that he could have exercised self-control by refraining from eating chocolate cannot be that he normally succeeds in exercising self-control on such occasions when he tries. And, in that case, when we say of John that he could have exercised self-control by refraining from eating the chocolate on that occasion, it cannot be that that is what makes what we say of him true either.

Honoré thus seems to me to be wrong that the sense of "can" in which we are interested—the sense in which John can exercise self-control, though he might fail to do so—is the sense of "can" (general). But if Honoré is wrong, then we might begin to wonder whether any coherent sense can be made of the idea that John can exercise self-control, and yet fail to do so, at all. Perhaps we should conclude instead that our commonsense assumption that there is a difference between an agent who has, but fails to exercise, a capacity for self-control, and an agent who has no capacity for self-control at all, is simply mistaken, and revise our allocations of moral and legal responsibility accordingly. However this would, I think, be premature. The sense of "can" can coherently be spelt out (Smith, 1997).

We want to know what the difference is between, on the one hand, the possible agent who has the capacity to exercise self-control by performing an action, but fails to exercise that capacity, and, on the other hand, the possible agent who has no such capacity, and who therefore fails to perform the action

because he could not have performed it. My suggestion is that the difference between these two agents consists in the differential similarity relations that obtain between the possible worlds in which they each fail and the possible worlds in which they succeed. Roughly speaking, the first of the two possible agents possesses an intrinsic feature which makes the possible world in which he succeeds in performing the action *more similar*, relatively speaking, to the possible world in which he fails, than the possible world in which the second of the two possible agents succeeds in performing the action is to the possible world in which he fails. This intrinsic feature possessed by the first agent is in turn, I suggest, what his capacity for self-control consists in. For this is what explains why it is possible (relatively speaking) for him to succeed in exercising self-control, whereas it is not possible (relatively speaking) for the second of two agents to succeed in exercising self-control (Smith, forthcoming).

Note that this suggestion explains why the capacity to exercise self-control comes in degrees. For two agents may be alike in that they both possess the capacity to exercise self-control and yet differ in that the one may find it easier to exercise self-control than the other. Suppose we fix on a possible world in which three agents fail to perform some action. It might be true of two of them that they possess an intrinsic feature which makes the possible world in which they succeed in performing the action in question more similar, relatively speaking, to the possible world in which they fail, than is the possible world in which the third agent succeeds to the possible world in which he fails. Yet it might also be the case that the possible world in which one of the two succeeds is more similar to the world in which he fails, relatively speaking, than the possible world in which the other succeeds is to the possible world in which he fails. This is why sanction and blame come in degrees as well. It would plainly be unfair to sanction or blame, to the same extent, two people who differ in the crucial respect that one of them is more like someone who shouldn't be sanctioned or blamed at all.

The attraction of this suggestion should be plain. The difficulty is to explain what makes claims about our capacities true or false, and the solution is to suppose that what makes claims about our capacities true or false is exactly the same sort of thing that makes any other modal claim true or false: namely, facts about the similarities that obtain between possible worlds. The idea of an unexercised capacity for self-control, as opposed to a lack of self-control, therefore turns out to be no more mysterious than the idea of a possible world in which there is something that is really quite similar to the way that that very thing is in another possible world, as opposed to a possible world in which there is something that is not similar at all to the way that that very thing is in another possible world.

I said above that I had some sympathy with Honoré's suggestion that to say of someone that he has, but fails to exercise, a capacity for self-control is to say of him that he usually succeeds in exercising self-control. The reason that I am sympathetic should now be clear. If the modal facts I have described are what makes claims about our abilities true or false, then it comes as no surprise that the best evidence that we have for the truth of some particular claim to the effect that someone or other can exercise self-control, when he fails to do so, will often be that that person usually does succeed in exercising self-control when he tries to do so. Regular patterns in actuality are, quite in general, what provide us with such evidence as we have for the similarities and differences that obtain between the actual world and other possible worlds. But we must not let this epistemological point obscure the metaphysics of capacities. What makes claims about our capacities true or false are the modal facts I have described, not the regular patterns in actuality that would provide us with evidence of those modal facts.

Here, then, lies the solution to the fifth puzzle about self-control. The difference between an agent who has, but fails to exercise, a capacity for self-control and another agent who has no capacity for self-control at all, lies in the relative nearness or remoteness of the possible worlds in which such agents succeed in exercising self-control from the possible worlds in which they fail to exercise self-control. To repeat, the best evidence for the nearness or remoteness of such possible worlds lies, much as Honoré suggests, in whether or not the agents in question usually succeed in exercising self-control when they try. But this claim about our evidence for the truth of ascriptions of the capacity for self-control must not be offered as a substitute for what makes such ascriptions true.

Conclusion

I said at the outset that my aim was to raise, and hopefully to answer, some of the difficult questions that arise given that we restrict sanctions and blame to those who have rational control over their conduct. The main conclusions can be summed up as follows.

Though possession of an intention suffices for agents to be in control 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 intentions are suitably responsive to their deliberations, that is, to their reflectively formed beliefs about what they would want themselves to do if they were fully rational. The capacity for self-control can thus be seen to embody

this responsiveness. It is the capacity rational agents possess to have desires corresponding to those they believe they would have if they were fully rational, 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 rational self-control we can define the idea of someone who, though capable of acting intentionally, remains a victim of circumstance. An agent who, though capable of acting intentionally, remains a victim of circumstance is someone who, on the one hand, has the capacity to act on his desires and intentions, but is also someone who, on the other hand, has desires and intentions that are beyond the reach of the capacity he has for rational self-control. The limits of an agent's capacity for self-control is in turn fixed by the nearness or remoteness of the possible worlds in which he succeeds in exercising self-control from actuality. This explains why the capacity for rational self-control comes in degrees.

The distinction between synchronic and diachronic exercises of self-control is, however, crucial at this point. For an agent to truly lack self-control, and hence to be truly a victim of circumstance, more must be true of him than that he has a desire which is he is unable to conquer synchronically. In other words, more must be true than that no feat of the imagination or thought which was within his reach at the time could have stopped the desire from having its effect. The desire must also be one which the agent could not reasonably have foreseen that he would have at a time at which he was not out of control. 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 have a desire that is beyond the reach of his capacity for synchronic self-control, does not have a desire that is beyond the reach of his capacity for diachronic self-control. Such an agent is thus not a victim of circumstance, notwithstanding his inability to exercise synchronic self-control.

References

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

Davidson, Donald (1970) "How is Weakness of the Will Possible?" reprinted in his *Essays on Actions and Events* (Oxford: Oxford University Press, 1980), 21–42.

Frankfurt, Harry (1971) "Freedom of the Will and the Concept of a Person" reprinted in Gary Watson (ed.), *Free Will* (Oxford: Oxford University Press, 1982), 81–95

Hobbes, Thomas (1651) *Leviathan* (Harmondsworth: Penguin, 1968).

Honoré, Tony (1964) "Can and Can't" reprinted as "Appendix: Can and Can't" in his *Responsibility and Fault* (Oxford: Hart Publishing, 1999), 143–60.

——(1998) "Being Responsible and Being a Victim of Circumstance" reprinted in his *Responsibility and Fault*, 121–42.

Kennett, Jeanette and Michael Smith 1996. "Frog and Toad Lose Control", *Analysis* 56, 63–73.

Pettit, Philip and Michael Smith (1996) "Freedom in Belief and Desire", *Journal of Philosophy* 93, 429–49.

Smith, Michael (1994) *The Moral Problem* (Oxford: Basil Blackwell).

——(1995) "Internal Reasons", *Philosophy and Phenomenological Research* 55, 109–31.

——(1997) "A Theory of Freedom and Responsibility" in Garrett Cullity and Berys Gaut (eds), *Ethics and Practical Reason* (Oxford: Oxford University Press), 293–319.

——(2000) "Quelques énigmes concernant le contrôle de soi", *Philosophiques* (27), 287–304.

——(forthcoming) "Rational Capacities", in manuscript.

Watson, Gary (1975) "Free Agency" reprinted in Gary Watson (ed.), *Free Will*, 96–110.