Meeting of the Aristotelian Society held at Senate House, University of London, on 18 *June* 2012 *at* 4:15 *p.m.*

XV—Agents and Patients, or: What We Learn About Reasons for Action by Reflecting on Our Choices in Process-of-Thought Cases

Michael Smith

Can we draw substantive conclusions about the reasons for action agents have from premisses about the desires of their idealized counterparts? The answer is that we can. The argument for this conclusion is Rawlsian in spirit, focusing on the choices that our idealized counterparts must make simply in virtue of being ideal, and inferring from these choices the contents of the desires that they must have. It turns out that our idealized counterparts must have desires in which we ourselves figure as both agents and patients, and in which others must figure too, though only as patients.

Can we draw substantive conclusions about the reasons for action agents have from premisses about the desires of their idealized counterparts? Many will agree that we can if we stipulate a connection between reasons for action, on the one hand, and idealized desires, where the norms governing the idealization are moral norms, on the other. But what if the ordinary concept of a reason for action is in play, and hence that the norms governing the idealization are norms internal to the concept of agency itself? For example, what if we think of an agent as a functional kind, defined by the possession and exercise, to some degree or other, of the capacities to know the world in which he lives and realize his desires in it, and hence think of the norms governing the idealization as those to which an agent conforms when he fully and robustly possesses and exercises this pair of capacities? My own view is that substantive conclusions follow from premisses like these, premisses that make no moralized assumptions about the norms internal to the concept of agency (see also Smith 1994).

In saying this I am, of course, swimming against a tide. Bernard Williams (1981) famously argues that an agent has a reason to act

Proceedings of the Aristotelian Society, Vol. cxii, *Part* 3
doi: 10.1111/j.1467-9264.2012.00337.x

in a certain way in certain circumstances, in the ordinary sense, just in case he would desire that he acts in that way in those circumstances if he were to deliberate correctly, and he further argues that correct deliberation has to be understood in terms of an agent's full possession and exercise of the two capacities that I have said are internal to the concept of an agent. Williams's account of the ordinary concept of a reason for action is thus much the same as my own, but he thinks that substantive facts about what agents have reason to do, in this sense, are all relative to what their potentially idiosyncratic and immoral desires happen to be, a conclusion he takes great delight in emphasizing (see, for example, his discussion of the reasons for action that a cruel husband might have: Williams 1995). There is therefore confusion, or so it might seem, in my supposing that we could move from Williams-style anti-rationalist premisses about the nature of reasons for action to a rationalist conclusion about the substance of the reasons that people have. But though there is some truth to this characterization of my argument as moving from anti-rationalist premisses to a rationalist conclusion, I will argue that it involves no confusion. As I see things, theorists like Williams should abandon their anti-rationalism for reasons internal to their own understanding of what it is to deliberate correctly.

Though the core of the argument I will go on to give is rather different from those given by other rationalists for similar conclusions, it is not unlike the Original Position argument Rawls gives in *A Theory of Justice* (1971). Rawls invites us to imagine ourselves choosing principles to govern the basic structure of a society in which we will eventually live, but he tells us that we must make this choice from a position in which we are ignorant about who we will eventually turn out to be in that society, and ignorant of what our potentially idiosyncratic desires are when we make that choice. Since anti-rationalists think that all choices express our potentially idiosyncratic desires, they must find Rawls's suggestion that we can make such a choice utterly baffling. One attraction of Rawls's argument, as I see things, is thus that it focuses our attention on what's really at issue in the debate between rationalists and anti-rationalists, namely, the possibility of a choice made in circumstances like those he describes.

If a choice of the kind that Rawls imagines us making in the Original Position is so much as possible, as both he and I think it is, then that choice must itself be the expression of a desire we have to have

doi: 10.1111/j.1467-9264.2012.00337.x

simply in virtue of being capable of rational choice (we have, after all, abstracted away from our potentially idiosyncratic desires, so all that's left to ground that choice is a desire we have to have simply in virtue of being rational choosers), and the content of that desire mustn't have anything especially to do with ourselves (as we have also imagined ourselves ignorant of who we are, so all self-concern would be idle). The coherence of Rawls's imagined choice thus augurs in favour of rationalism and against anti-rationalism. The argument I go on to give is a lot like Rawls's in that it focuses on choices made in circumstances that are, in crucial respects, very similar to those that he imagines in the Original Position. Moreover, as we will see, the substantive upshot is also very similar to Rawls's. But it is unlike Rawls's argument in focusing on much more mundane choices. The advantage of attending to more mundane choices is that it turns out to be far less controversial that we would in fact make these choices in the imagined circumstances. The argument is thus an improvement on Rawls's, or so it seems to me.

I have said that theorists like Williams should abandon their anti-rationalism for reasons internal to their own conception of correct deliberation. So what is wrong with that conception? The problem is that the conception purports to be one that honours requirements of coherence, but fails spectacularly to do so. The psychology of an agent who fully and robustly possesses and exercises the capacity to have knowledge of the world in which he lives and realize his desires in it is supposed to be one that, among other things, realizes the virtues of coherence in both the theoretical and practical domains. The evidence available to the agent is supposed to cohere with the beliefs he forms on the basis of that evidence, and his non-instrumental desires, his beliefs about means, his instrumental desires and his actions are also supposed to cohere with each other in familiar ways. But, as we will see, exercises of these two capacities seem not to cohere at all well with each other. That's the problem.

To see the lack of coherence, remember that the capacity for desire-realization is one that an idealized agent is supposed to have no matter what the content of his desires turns out to be. His desires can be utterly idiosyncratic, perhaps even bearing on the exercise of his capacity to believe for reasons, or the exercise of his capacity to realize his desires. In this spirit, let's think more about the former case. Imagine an agent who desires that he now believes that *p*, but imagine further that *p* isn't the case and that the evidence that *p* isn't

doi: 10.1111/j.1467-9264.2012.00337.x

the case is available to him. This agent is in a synchronic bind. If he fully and robustly exercises his capacity for desire-realization, he cannot fully and robustly exercise his capacity for belief-formation, and vice versa. Nor would he have been any better off if he had had no such desire. He would have been no better off because the coherence of the deliverances of the two capacities would at best have been just a happy accident. The mere fact that there exist possible worlds in which the deliverances of the two capacities diverge is thus sufficient to make it impossible for agents fully *and robustly* to exercise the two capacities that Williams tells us underwrite correct deliberation.

If we were to stick with Williams's conception of an ideal psychology, we would be reduced to giving separate scores for the extent to which an agent's psychology is ideal along the different dimensions. (By way of analogy, think of how in Olympic diving separate scores are given for the difficulty of a dive and its execution.) There would be one score for the extent to which an agent possesses and exercises the capacity for accessing evidence and forming beliefs in the light of that evidence, and a separate score for the extent to which he possesses and exercises the capacity for desire-realization, making his instrumental desires cohere with his non-instrumental desires and means–end beliefs, and acting accordingly. The question is whether we can rest content with this separate scores conception of an ideal psychology, and the answer turns on whether an alternative conception of an ideal psychology is available, one that makes the two capacities cohere better with each other. If so, then given that the norms governing an ideal psychology give pride of place to norms of coherence, we would do better to adopt that alternative conception of an ideal psychology.

There are at least three ways in which we could revise our conception of an ideal psychology so as to ensure that that psychology is more robustly coherent. The first would be to suppose that the coherence of an agent's psychology is wholly determined by the extent to which he possesses and exercises the capacity to realize his desires. The possession and exercise of the capacity to know his world might still be required for having an ideal psychology from time to time, but only in so far as it contributes to desire-realization. The second would be for the coherence of an agent's psychology to be wholly determined by the extent to which he possesses and exercises the capacity to access evidence and form beliefs on its basis. The

doi: 10.1111/j.1467-9264.2012.00337.x

possession and exercise of the capacity to realize desires might still be required for having an ideal psychology from time to time, but only in so far as it contributes to knowing his world. On both of these ways of revising our conception of an ideal psychology, the separate scores conception is abandoned because we deny that the standard that generates one of the scores is a standard at all.

Both of these ways of revising our conception of an ideal psychology are unacceptably revisionary. The first way of revising our conception of an ideal psychology admits that such a psychology is prone to all sorts of dysfunction in the formation of beliefs, but then pretends that that dysfunction doesn't amount to incoherence. The second way of revising our conception of an ideal psychology tells us that such a psychology is prone to all sorts of dysfunction in the formation of instrumental desires, but then pretends that that dysfunction doesn't amount to incoherence. The separate scores conception is preferable to each of these because it at least acknowledges that dysfunction in the formation of beliefs and instrumental desires is as such a departure from a norm internal to the concept of agency, and thus amounts to incoherence within a psychology.

There is, however, a third way in which we could revise our conception of an idealized psychology, a way that also admits that dysfunction in the formation of beliefs and instrumental desires is a departure from a norm internal to the concept of agency, but which is far less revisionary. According to this third way, having certain coherence-inducing desires is partially constitutive of what it is to have an ideal psychology. Focus again on the agent who desires that he now believes that *p*. If in order to have an ideal psychology, that agent has to have a dominant desire that he does not now interfere with his current exercise of his belief-forming capacities, where a dominant desire is one that overrides all of the potentially idiosyncratic desires he has—that is, the desires that aren't partially constitutive of his being ideal—then there is no conflict of the kind we have been worrying about between the full and robust exercise of an agent's belief-forming capacities and the full and robust exercise of his desire-realization capacities. The only way in which he could possess and exercise the latter in worlds in which he is otherwise ideal, but desires to believe that *p*, would be by leaving himself free to exercise the former.

This third way of revising our conception of an ideal psychology is therefore preferable to the other two because it doesn't require us

doi: 10.1111/j.1467-9264.2012.00337.x

to suppose that dysfunction in the formation of beliefs or instrumental desires is an integral part of the ideal. It admits that dysfunction as such amounts to incoherence. Greater coherence is achieved not by pretending that no real coherence was ever to be gained by an agent's exercising one or another of his capacities to believe for reasons or realize desires, but is rather achieved by insisting that an ideal agent, one who fully and robustly possesses and exercises both of these capacities, has a dominant coherence-inducing desire to not now interfere with his current exercise of his belief-forming capacities. This desire is coherence-inducing because it ensures that his exercise of his desire-realization capacities chimes with his exercise of his belief-forming capacities.

I said at the beginning that Williams takes delight in emphasizing that an ideal agent's desires can be as idiosyncratic and immoral as you like. But if the conclusion just argued for is correct, Williams is mistaken. Though for all we've said so far an ideal agent may well have desires that are idiosyncratic, and perhaps even immoral, these desires will be dominated by a desire that an ideal agent has to have, simply in virtue of being ideal. All ideal agents have to have a desire that they do not now interfere with their current exercise of their capacity to believe for reasons. Moreover, the argument given for this conclusion has been, just as I said it would be, thoroughly Rawlsian in spirit. We discover that an ideal agent has to have the coherence-inducing desire we have identified by reflecting on a choice that he can and must make, in so far as he is ideal, a choice that no idiosyncratic desire he has could possibly explain.

When we put this conclusion together with the idea that the reasons for action that we have are fixed by the desires of our idealized counterparts, the upshot is that all agents have a reason not to interfere with their own current exercise of their belief-formation capacities, no matter what their idiosyncratic desires happen to be, and that they have this reason in virtue of the coherence-inducing desire that their idealized counterparts have to have simply in virtue of being ideal. The main claim of the paper has therefore been established. Premisses about the desires of agents' idealized counterparts do indeed entail substantive conclusions about their reasons for action. But the argument itself suggests that we shouldn't stop at this point. For once we see that ideal agents have to have one dominant coherence-inducing desire, and hence that all agents have one substantive reason for action in common, an obvious question to ask is

doi: 10.1111/j.1467-9264.2012.00337.x

whether they have to have any other such desires, and hence whether they have any other substantive reasons for action in common. Of ultimate interest, of course, is whether any such reasons have recognizably moral content.

So far we have focused on what it takes for an agent fully and robustly to possess and exercise the capacity to believe for reasons and realize his desires. But given that agents can exist over time, and given that the formation of many of their desires and beliefs therefore takes time, there are at least two very different ways for an agent to meet this general description. He might fully and robustly possess and exercise the capacity to believe for reasons and realize his desires at the present moment, but without regard for what's happening at any of the future moments at which he exists. Alternatively, he might fully and robustly possess and exercise the capacity to believe for reasons and realize his desires at the present moment, but in such a way as to make sure that it is possible for himself to fully and robustly possess and exercise the capacity to believe for reasons and realize his desires at the future moments he exists as well, at least to the extent that he can have an effect on what's happening in the future at the present moment. Which of these two ways of being an agent is more ideal? Since being ideal includes, at a minimum, being consistent, in the sense of treating like cases alike, it seems that the agent who fully and robustly possesses and exercises these capacities not just at the present moment, but in such a way that he can do the same thing at later moments, is plainly more ideal. He is more ideal because he is more consistent.

Imagine an agent who is in the business of exercising his capacity to believe for reasons by engaging in a process of thought that takes time. Right now he is trying to figure out whether *p* is true. If it turns out to be true, then he will later try to figure out whether *p* supports *q*. If it does then, relying on his memory of having established that *p* and that *p* supports *q*, he will draw the conclusion that *q*. However let's also imagine that he now has the desire to believe that not *q* later. If this agent fully and robustly possesses and exercises the capacity to believe for reasons and realize his desires now, then even if he has the required desire that he does not now interfere with his current exercise of his belief-forming capacities, he would still be in a similar situation, later, to the synchronic bind we've been talking about. In those worlds in which *q* is true, and the evidence that this is so is available to him, if he is now to succeed in fully and robustly ex-

doi: 10.1111/j.1467-9264.2012.00337.x

ercising his capacity to realize his desires, he would later have to end up believing that not *q*. His present exercise of his capacity to realize his desires would therefore have to come at the cost of his later failure to exercise his capacity to believe for reasons.

Note that the cause of this problem lies fairly and squarely with the agent's having a desire now with a content that is similar to the content of the desire responsible for the synchronic bind we've been talking about. The cause of the synchronic bind was the agent's desire to believe that *p now*; the cause of this problem is the agent's desire to believe that *p later.* To the extent that the agent who has the desire we have already seen to be partially constitutive of his being ideal can address the problem caused by the former desire, but cannot address the problem caused by latter desire, he seems to be inconsistent, in the sense of not treating like cases alike. So what's required for him to be consistent? What's required is that he has another desire that dominates the latter desire. In other words, to be consistent in his treatment of these two cases, the agent must have an additional dominant desire, a desire that he does not now interfere with his later exercise of his belief-forming capacities. The possession of this further desire must also be partially constitutive of an agent's being ideal.

A similar line of reasoning suggests that further dominant desires are required as well. Imagine the same case, but instead of supposing that the agent now has a desire to believe that not *q* later, suppose that he is feeling slightly distracted. Perhaps feeling distracted is a condition he has from time to time, one that comes in waves when he has low blood sugar, something that varies predictably during the day. His feelings of distraction make it hard for him to concentrate, but not so hard as to interfere with his exercise of his capacity to figure out whether *p* is true. He has one sugar pill that would get rid of his feelings of distraction, so making it easier for him to think now. However he also knows that, even if he gets rid of his feelings of distraction now, the feelings of distraction will return later when, if all goes well, he will be trying to figure out whether *p* supports *q*, and that he will then feel so distracted that his capacity to figure out whether *p* supports *q* will be severely diminished. The question is whether to take the sugar pill now, or to hold on to it and take it later.

It seems to me that if the agent we are imagining is ideal, in the sense that he fully and robustly possesses and exercises the capacity

doi: 10.1111/j.1467-9264.2012.00337.x

to form beliefs for reasons and realize his desires not just in the present, but in the present in such a way as to make it possible for himself to fully and robustly possess and exercise the capacity to believe for reasons and realize his desires in the future, then we already know the answer to this question. His present indifference to his future possession of the capacity to form beliefs for reasons is what sets him up so as not to possess and exercise that capacity later. So if the agent is ideal, he must have a desire to hold onto the sugar pill and take it later. But nothing said so far guarantees that an ideal agent does have this desire. Since the feelings of distraction will diminish his future capacity to believe for reasons, it needn't be the case that his future exercise of this capacity is being interfered with by his not holding on to the sugar pill now so that he can take it later. The only conclusion to draw is that the ideal agent must have yet another dominant desire as well, a desire with a rather different content. In so far as he is ideal, he must desire that he now does what he can to help himself to later have the capacity to believe for reasons. His possession of this additional dominant desire, together with the fact that he fully and robustly exercises his capacity to realize his desires now, explains why, as an agent who isn't just ideal in the present, but is ideal in the present in such a way as to be ideal in the future as well, he would hold onto the sugar pill and take it later.

So far the focus has been entirely on what's required for an agent to possess and exercise the capacity for belief-formation. But nearly everything that has been said on this front applies, *mutatis mutandis*, to an ideal agent's desires concerning his possession and exercise of his desire-realization capacities as well. Agents who fully and robustly possess the capacity to realize their desires must be invulnerable not just to their having idiosyncratic desires concerning what to believe, but also to their having idiosyncratic desires concerning which desires to realize. Imagine, for example, an agent who now knows that he will later desire that *p*, and who now, desiring that he later realizes that desire to the exclusion of others, lays traps for his future self so as to ensure that no other desire that he might happen to have later is realized instead. Perhaps he now makes sure that he is later instrumentally irrational with respect to every other desire except the desire that *p*, or perhaps he now simply ensures that the only option he will have later is the option of satisfying the desire that *p*. Agents who fully and robustly possess and exercise the capacity to realize

doi: 10.1111/j.1467-9264.2012.00337.x

their desires not just in the present, but in the present in such a way as to do so in the future as well, do not lay such traps for themselves (compare the discussion of prudence in Nagel 1970).

An ideal agent must therefore have versions of the desires already described that concern his exercise of his belief-formation capacities, but these desires must concern his exercise of his desire-realization capacities. He must have a dominant desire that he does not now interfere with his later exercise of his capacity to realize his desires, and he must also have a dominant desire that he now does what he can to help himself to later have the capacity to realize desires. These desires are, however, subject to a crucial proviso. The desires with which the ideal agent desires not to interfere must themselves be desires whose realization doesn't require that he interferes with the exercise of his capacities for belief-formation or desire-realization. But since this proviso is so obvious and such a mouthful, I will take it as read in everything that follows.

I have argued that an ideal agent fully and robustly possesses and exercises two capacities in the present in such a way that he can do the same thing in the future—the capacity to access evidence and form beliefs on the basis of that evidence, and the capacity to realize his desires—and I have further argued that if his exercises of these two capacities are to cohere with each other, then he must have a whole slew of dominant coherence-inducing desires: the desires that he does not now interfere with his exercise of his capacity to believe for reasons or realize his desires either now or later, and the desires that he now does what he can to help his later self have belief-formation and desire-realization capacities. These desires must all be dominant in the sense that they must dominate all of the other potentially idiosyncratic desires that he might happen to have. But how strong are these dominant desires supposed to be vis-à-vis each other? Which should dominate which when they come into conflict?

For example, imagine an ideal agent who is suffering from a disease that will eventually cause him to lose all of his mental powers, absent treatment, and further imagine that the disease is especially susceptible to mood, so that the treatment regime includes a pill that will cause sufferers to believe that they will get better. The ideal agent's dominant desire that he now does what he can to help his later self have belief-formation and desire-realization capacities will tell in favour of his taking the pill, but his dominant desire that he does not now interfere with his current exercise of his capacity to

doi: 10.1111/j.1467-9264.2012.00337.x

believe for reasons will tell in favour of his not taking it. His dominant desires therefore conflict. How would such a conflict be resolved in an ideal agent? Which of his desires, each of which must dominate his idiosyncratic desires, should dominate which?

Though I take it to be obvious that in this case the desire to help his later self have belief-formation capacities should dominate, if we are to give this answer on principled grounds then we would have to consider a whole range of conflict cases, so that we can see how it follows from a comprehensive account of the relative strengths of the different desires an ideal agent has to have vis-à-vis each other in different circumstances. Given that such an account might well be highly circumstance-specific, I will not even attempt to provide it here. However I mention it just to make it clear that even an ideal agent may find himself in circumstances in which the exercise of his capacity to believe for reasons comes at the cost of his exercise of his capacity to realize his desires, and vice versa, and hence that there will be principled limits to the robust exercise of these two capacities.

Does this mean that, at the end of the day, the conception of the ideal agent argued for here is no better than the separate scores conception? No, it does not mean that. According to the conception of an ideal agent argued for here, the conflicts that the ideal agent experiences are themselves all conflicts between desires that an agent has to have simply in virtue of being ideal, and such conflicts can therefore be resolved in a principled way, specifically, by reference to the relative strengths that these desires have to have vis-à-vis each other simply in virtue of being the desires of an ideal agent. This is the antithesis of the separate scores conception. What these residual conflicts show is not that the conception of an ideal agent argued for here is no better than the separate scores conception, but rather that, though the conception argued for here is better than that conception, as it secures more coherence in the psychology of an ideal agent than that conception does, there are principled limits to how much coherence there can be. The dominant desires constitutive of being ideal cannot always be co-satisfied.

Let's take stock. What's especially striking about all of the desires we have described so far, as candidate desires of an ideal agent, is that the agent himself occurs twice over in their content. He occurs in both the *agent*-place in these desires ('… the agent desires that *he* does not now interfere with …', '… the agent desires that *he* now does what he can to help …') and in the *patient*-place ('… interfere

doi: 10.1111/j.1467-9264.2012.00337.x

with *his* exercise of *his* capacities now or later …', '… help *his* later self have the capacity …'). However, he only occurs in the agent-place in these desires in the present ('… the agent desires that he does not *now* interfere with …', '… the agent desires that he *now* does what he can to help …'), whereas he occurs in the patient-place in both the present and the future ('… interfere with his exercise of his capacities *now or later …*', '… help his *later* self have the capacity …'). Other people, by contrast, don't occur in the content of these desires at all, not in the agent-place, and not in the patient-place either. Is distinguishing in this way between the different ways in which the agent and others figure in the content of the desires ideal agents have to have, simply in virtue of being ideal, defensible?

It is clearly not just defensible, but essential that *the agent himself*, rather than other people, occurs in the agent-place in the desires that he has to have simply in virtue of being ideal. The role of these desires is to ensure that the agent himself fully and robustly exercises his capacities for belief-formation and desire-realization, not that other people exercise these capacities on his behalf, whatever that might mean, or that other people exercise their own capacities. To play this role, an agent's desires must therefore be desires about what he is to do, and, moreover, they must also be desires about what he is to do *at the present moment*. All desire-realization is, after all, initiated in the here and now. The agent himself at the present moment must therefore occur in the agent-place of the desires that an ideal agent has to have, simply in virtue of being ideal.

It is also clearly essential that the agent *both in the present and in the future* occurs in the patient-place of the desires that an ideal agent has to have simply in virtue of being ideal. This is because the role of these desires isn't just to ensure that the agent fully and robustly exercises his capacities for belief-formation and desire-realization at the present moment, but is also to ensure that he does so in such a way as to make it possible for himself to possess and robustly exercise those capacities in the future as well. The desires an agent has to have, simply in virtue of being ideal, must therefore concern what he himself is to do at the present moment that will have an effect on himself both in the present and in the future.

But what about *other people*? Should they occur in the patient-place of the desires that an ideal agent has to have simply in virtue of being ideal? Recall that our reason for thinking that the agent himself at later times has to occur in the patient-place was that it

doi: 10.1111/j.1467-9264.2012.00337.x

would be inconsistent for him to treat his later self differently from the way in which he treats his present self. It would be inconsistent for him to desire not to (say) now interfere with his own current exercise of his capacity to believe for reasons, but to be indifferent towards his now interfering with his later exercise of his capacity to believe for reasons. We must therefore ask whether it would be similarly inconsistent for an agent to treat other people differently from the way in which he treats his current or later self. Would it be inconsistent for an agent to desire not to (say) now interfere with his own current or future exercise of his capacity to believe for reasons, but to be indifferent towards his now interfering with other people's exercises of their capacity to believe for reasons? If so, then other people would have to occur in the patient-place of the desires that agents have to have simply in virtue of being ideal. We would be well on the way to showing that everyone shares reasons for action that have recognizably moral content.

There are various ways in which we could try to answer this question. One would be to argue on metaphysical grounds, in the manner of Derek Parfit's argument against the Self-Interest Theory in *Reasons and Persons*, that times and agents have to be treated similarly in the statement of principles that govern rational conduct (see Parfit 1984, especially §55). An agent's rational concern, whatever its substantive content, must therefore be either restricted to himself in the present or extended, not just to himself at later times, but also to other agents. This would in effect be to give a strengthened version of Parfit's argument against the Self-Interest Theory in *Reasons and Persons*. The strengthening of the argument derives from the fact that, since we have already seen that an ideal agent has to have desires that concern himself at other times, the only alternative left is to suppose that he also has to have desires that concern other agents too (see also Pettit and Smith 1997).

Another way to approach this question would be to ask whether it is conceptually possible for there to be a multitude of agents who all fully and robustly exercise their capacities for belief-formation and desire-realization. If the answer is that yes, this is a conceptual possibility, then we could infer that other agents must occur in the patient-place of the desires that ideal agents have to have, simply in virtue of being ideal, whether or not there exist other such agents. This is because if they didn't occur in the patient-place, then in those possible worlds in which other ideal agents exist and interact with

doi: 10.1111/j.1467-9264.2012.00337.x

each other, they would each be vulnerable to others' interfering with their exercises of their capacities for belief-formation and desire-realization, and for giving them no help in their having belief-formation and desire-realization capacities. Since such interference and lack of help would undermine the *robustness* of their exercises of their capacities for belief-formation and desire-realization, it follows that the very idea of a multitude of ideal agents for whom interaction is possible brings with it the idea of each such agent being able to rely on others, if there are any others, not to interfere and to help (see also Smith 2011).

But even if arguments like these do show that others must occur in the patient-place of the desires ideal agents have to have, simply in virtue of being ideal, they fall short of making the content of these desires sufficiently determinate. For example, one question to which we need an answer is whether the desires ideal agents have to have not to interfere with each others' exercises of their capacities to believe for reasons are restricted to those agents whom they might interfere with having the capacity to be ideal. If these desires are restricted in this way, then that would severely limit the class of beings whose belief-formation capacities we have reasons not to interfere with, as only those who have a global capacity to know their world and realize their desires in it would be the proper objects of such reasons.

This is just an example. The more general question is whether any of the desires ideal agents have to have, simply in virtue of being ideal, are restricted in any way that limits the class of beings we have reasons to affect. To answer this question, we once again need to consider arguments more like Rawls's. Let's therefore return to the process-of-thought cases we have been focusing on, but let's ask a slightly different question about them, a question that is even more like the question Rawls asks about our choice in the Original Position.

Imagine a subject who is totally engrossed in a curiosity-driven process of thought, following an argument wherever it leads him. At the present moment his attention is entirely focused on trying to figure out whether *p* is true, so much so that he is oblivious not just to his surroundings, but also to the other aspects of his mental life. If *p* turns out to be true, then the subject's attention will spontaneously shift to figuring out whether *p* supports *q*. If he concludes that it does, he will attend to this fact, put it together with the fact that *p*, and go on to draw the conclusion that *q*. Imagine that everything

doi: 10.1111/j.1467-9264.2012.00337.x

works out, and that the subject draws the conclusion that *q*. There is therefore a stream of consciousness in which the subject at different moments attends to the different parts of a problem, indeed, the very same problem as that to which the agents have attended in the other process-of-thought cases we have so far discussed.

Now focus on the subject at the moment at which he is trying to figure out whether *p* is true, and imagine that he knows one further fact about what's going on. Though some of the subjects engaged in this curiosity-driven process-of-thought case are identical to himself, not all of them are. What must this agent be like, at the moment at which he is wondering whether *p* is true, if he is fully and robustly to possess and exercise the capacity for belief-formation, given that he has this extra bit of knowledge? We already know the answer in schematic terms. He must have certain coherence-inducing desires concerning himself in the present and future, and consistency requires him to have those same desires concerning others. But since we're trying to be more precise about the features that agents have to have in order objects of this concern, let's rehearse the arguments one more time, but this time making as few assumptions as possible about the features agents have to have in order to be involved in such a process of thought. Let's not even make any assumptions about what's required for the identity of a subject over time. Instead, let's consider what we would say about this case in the light of a range of standard views about personal identity, and let's see what lessons we learn.

To begin, then, let's suppose that the Cartesian soul view is the correct account of personal identity. Our question is what the Cartesian soul who is wondering whether *p* is true would have to be like in order to fully and robustly possess and exercise the capacity to believe for reasons at that moment, on the assumption that he knows that his soul will persist and so be engaged in some of the rest of the process of thought, but not all of it. Since this case is the same as those we have already discussed, the answer has to be the same as before. The Cartesian soul must desire not to (say) interfere with his present or later exercises of his capacity to believe for reasons, in so far as he persists, and consistency demands of him that he desires that he does not now interfere with the exercises of the capacities to believe for reasons of any of the other Cartesian souls who might be engaged in that process of thought either. In other words, the knowledge that he will be involved in some, but not all,

doi: 10.1111/j.1467-9264.2012.00337.x

of that process doesn't have any effect on what he has to be like in order to fully and robustly possess and exercise the capacity to believe for reasons at that moment.

Now suppose that the Cartesian soul who is trying to figure whether *p* is true also knows that, though one of the other Cartesian souls who will eventually be involved in the process of thought has the capacity to play his role in that process, he doesn't have the global capacity to know his world and realize his desires in it. Would this have any effect on what the Cartesian soul who is trying to figure out whether *p* is true has to be like in order to fully and robustly possess and exercise the capacity to believe for reasons at that moment, so playing his role in that process of thought? The answer is that it seems to make no difference at all. To be more precise, in so far as he is ideal, the Cartesian soul who is trying to figure out whether *p* is true must desire that he does not now interfere with his present or later exercise of his capacity to believe for reasons, in so far as he persists, and he must in consistency have this desire with respect to the later Cartesian souls' exercises of their capacities to believe for reasons too; but all of these desires allow that those who are the objects of concern may not possess or exercise their capacities to believe for reasons or realize their desires fully or robustly. It is sufficient that that they have the capacity to believe for reasons to some extent. That's all that's required for a desire not to interfere with their exercise of their capacity to be warranted.

Now suppose that the bodily criterion is the correct account of personal identity, and let's fix on that part of the body that seems like the best candidate to be the part that secures identity over time, namely, the brain. The persistence of the subject of the process of thought is thus a matter of the persistence of his brain. Now imagine the same scenario again, except that, immediately after figuring out that *p* and that *p* supports *q*, the subject suffers a catastrophic accident that so damages his body that, in order to give him the best chance of survival, the doctors on the scene immediately bisect his brain and transplant the two halves into two de-brained bodies, bodies that bear no physical similarity to his body; that both transplants miraculously turn out to be successful; that the subject was so engrossed in the problem that he was trying to solve that there was no realization that either the accident or subsequent operations happened; and that the subject knew from the outset that he would at some point suffer such a catastrophic accident.

doi: 10.1111/j.1467-9264.2012.00337.x

In the imagined scenario, there are two streams of consciousness from the moment of bisection onwards, each continuous with the earlier stream, and in each of these streams a subject attends to the fact that *p* and that *p* supports *q* and draws the conclusion that *q*. But since no single brain underlies the process(es) of thought described, there is no identity of the initial subject over time. Even so, it seems plain that for the initial subject, the one who is trying to figure out whether *p* is the case, to fully and robustly possess and exercise the capacity to believe for reasons at that moment, so playing his role in the process of thought, he must desire that he does not now interfere with his own present or future exercise of his capacity to believe for reasons, and, in order to be consistent, he must also desire that he does not interfere with anyone else's exercise of their capacity to believe for reasons either, including the later products of brain bisection. What this case suggests is thus that, since knowledge that such a catastrophic accident would at some point occur wouldn't warrant withdrawal of the desire, it follows that those who fall within the scope of the desire not only need not possess or exercise their capacities to know their world and realize their desires in it fully or robustly, but that they need not have a body like the initial subject, or be physically similar to him, either. It is sufficient that they have the capacity to believe for reasons to some extent, whatever their physical embodiment.

What if we suppose that the identity of a subject over time is a matter of psychological continuity and connectedness between the different stages of that subject? In this case we have to imagine that, immediately after having figured out that *p* is the case and that *p* supports *q*, the bulk of the underlying psychology of the subject—all of the psychology except for the part that underwrites the process of thought itself—changes radically, and we further have to imagine that this is something that the agent knows is on the cards, and that even so he remains oblivious to its happening. Perhaps the underlying psychology is just like mine up until he has figured out that *p* supports *q*, and just like my wife's thereafter, or just like a child's, or just like that of someone suffering from dementia who has a moment of clarity and sees that *q* follows from the facts that *p* and that *p* supports *q*. Of course, there would still have to be some connections between the episodes of thought that constitute the process of thought itself for it to be a process of thought. The subject who asks himself whether *q*, given that *p* and that *p* supports *q*,

doi: 10.1111/j.1467-9264.2012.00337.x

must do so because he seems to remember having established that *p* and that *p* supports *q* in the past, and that in turn must be so because the earlier subject did indeed establish that *p* and that *p* supports *q*. But, relative to the psychology as a whole, these connections are plainly far too meagre to secure the identity of a single subject over time, given the criterion of psychological continuity and connectedness.

Even if all of this were the case, and even if the subject knew it to be on the cards, our judgement about what's required for the subject who initiates the process of thought to fully and robustly exercise his capacity to believe for reasons, so playing his role in the process of thought, remains unchanged. If the bundle of psychologically continuous and connected psychological states that includes the attempt to figure out whether *p* is the case is to include the full and robust possession and exercise of the capacity to form beliefs for reasons, then it seems that it must also include the desire not to interfere with an exercise of the capacity to believe for reasons that falls within that bundle itself, and it must also include the desire not to interfere with an exercise of the capacity to believe for reasons that falls within the bundle that tries to figure out whether *p* supports *q*; and then consistency requires it to include as well the desire not to interfere with the exercise of the capacity to believe for reasons that is included in the bundle of psychologically continuous and connected psychological states that includes the attempt to draw the conclusion that *q*.

What this case suggests is thus that the desires an ideal agent has to have, simply in virtue of being ideal, are desires that allow that those who fall within their scope not only need not possess or exercise their capacities for belief-formation or desire-realization fully or robustly; and need not have a body like his, or be physically similar to him; but that they need have virtually nothing in common with him psychologically either: they needn't be psychologically similar to him, nor need they have capacities for belief-formation that are anywhere near as sophisticated as his are. That's the lesson we learn from the fact that, in spelling out this variation on the case, the psychology surrounding the process of thought could be like mine initially, but like a child's, or that of someone suffering from dementia, later on. Identity, physical similarity, psychological similarity, and the possession of sophisticated belief-formation and desire-realization capacities, these are all excluded as candidate properties that

doi: 10.1111/j.1467-9264.2012.00337.x

characterize those beings who figure in the patient-place of the desires ideal agents have to have simply in virtue of being ideal.

Other features of these process-of-thought cases, though it has so far been helpful to keep them constant in order to see that certain other candidate properties are excluded, are also plainly irrelevant when it comes to characterizing the beings that figure in the patient-place of the desires ideal agents have to have simply in virtue of being ideal. For example, though in each of these process-of-thought cases we have imagined a stream of consciousness, it plainly isn't necessary that the agent who figures in the patient-place of an ideal agent's desires be conceived of as someone who is conscious. The desires in question therefore cannot be restricted to conscious agents. If an agent is fully and robustly to possess and exercise a capacity to believe for reasons not just in the present, but in the present in such a way that he can do so in the future as well, then he must desire that he does not now interfere with his later exercise of his capacity to form beliefs for reasons whether or not he is conscious at the later moment of interference. Nor is it necessary that the beings with whom he might interfere are engaged in the formation of beliefs whose contents are inferentially connected with the contents of the beliefs that he is currently entertaining with a view to making an inference. For, once again, if an ideal agent is fully and robustly to possess and exercise the capacity to believe for reasons, not just in the present, but in the present in such a way that he can do so in the future as well, then he must desire not to interfere with his later exercise of his capacity to form beliefs for reasons whether or not their contents are inferentially connected with the beliefs that he is currently entertaining.

So what features must beings possess if they are to figure in the patient-place of the desires that an ideal agent has to have, simply in virtue of being ideal? When we put what we have learned from our discussion of all these process-of-thought cases together, and generalize what we have learned from these cases to cases in which the exercise of desire-realization capacities is at issue as well, it seems that the only feature that beings who figure in the patient-place of the desires that ideal agents have to have, in so far as they are ideal, is that they have some rudimentary capacity to believe for reasons and realize their desires, and that they are such that their exercise of these capacities depends on what the agent now does. Dependency, together with a minimal capacity to form beliefs for reasons and re-

doi: 10.1111/j.1467-9264.2012.00337.x

alize desires, are the only features a being needs to have in order to figure in the patient-place of the desires the ideal agent has to have simply in virtue of being ideal.

The upshot is that we can characterize the psychology of the ideal agent as follows. The ideal agent must fully and robustly possess and exercise the capacity to access evidence and form beliefs on the basis of that evidence, and the capacity to realize his desires, both in the present and in such a way that he can do so in the future. This requires that he has certain coherence-inducing desires. In particular, it requires that he has a dominant desire that he does not now interfere with the exercise of the capacity to believe for reasons or realize desires of any being at any time whose exercise of their capacity to believe for reasons or realize their desires is dependent on what he now does, and it also requires that he has a dominant desire to do now what he can to help any such being at any time whose possession of belief-formation and desire-realization capacities is dependent on what he now does to have belief-formation and desire-realization capacities.

For short, let's call these the desires to help and not interfere. If the agent is ideal then the desires to help and not interfere must themselves be dominant in the sense that they must dominate any potentially idiosyncratic desires that the ideal agent might happen to have. Only so will the robust exercise of his two capacities be guaranteed. These desires may of course conflict with each other, but the ideal agent will resolve such conflicts by having each of these desires with an appropriate strength. I haven't said how strong these desires must be vis-à-vis each other, so there remains some unfinished business. But, unfinished business notwithstanding, it must surely be agreed that this represents considerable progress in answering the question we asked at outset.

Can we draw substantive conclusions about the reasons for action agents have from premisses about the desires of their idealized counterparts? The answer turns out to be that we can, and that these reasons for action have recognizably moral contents. Since the reasons for action agents have are a function of the desires of their idealized counterparts, and since every agent's idealized counterpart has to have the same dominant desires to help and not interfere alongside whatever other idiosyncratic desires he has, it turns out that every agent has dominant reasons to help and not interfere, and it also turns out that, beyond that, every agent has a reason to satisfy what-

doi: 10.1111/j.1467-9264.2012.00337.x

ever undominated idiosyncratic desires he happens to have. This is much like a standard liberal deontological view of our moral obligations and permissions, but with one important twist. Everyone with the capacity to have these coherence-inducing desires has a reason to make sure that not just people, but all beings with rudimentary capacities to know the world in which they live and realize their desires in it, have the wherewithal to lead a life in which they exercise these capacities, and hence a life that they, in a sense, choose. Moreover, everyone also has a reason to leave such beings free to lead lives of their own choosing, and to lead lives of their own choosing themselves. The twist should be clear. Human infants, mentally defective adult humans, and non-human animals, all of whom are typically problem cases for the standard liberal deontological view, are objects of our moral obligations and permissions from the outset according to the view defended here. This is because they are all equally such that their exercise of their capacities to know the world and realize their desires in it depend on what people with such obligations do.

As advertised, the argument for this conclusion has been thoroughly Rawlsian in spirit. It has focused on the choices that an agent has to be capable of making simply in virtue of being ideal, and it has inferred the contents of the desires that ideal agents have to have from the substantive choices that they make. But whereas Rawls asks us to imagine choosing principles that will govern the basic structure of a society in which we will live, and hence immediately courts controversy, the argument given here has focused on choices made in much more mundane circumstances in which our basic agential capacities are exercised. The focus throughout has been on choices made in process-of-thought cases, cases in which the choices to be made are much more clear-cut. The question we have asked is what an ideal agent would have to choose to do in order to bring his capacities for belief-formation and desire-realization into coherence with each other when he is engaged in certain processes of thought, and the answer has turned out to be that, quite generally, he would have to choose to help and not interfere.

But though the argument for this conclusion has focused on choices made in much more mundane circumstances than those Rawls asks us to imagine, the conclusion reached has turned out to be remarkably similar to his. The reasons we have to help and not interfere are, after all, not just similar in content to the standard lib-

doi: 10.1111/j.1467-9264.2012.00337.x

eral deontological view of our moral obligations and permissions, but are also more than somewhat reminiscent of Rawls's own second and first principles of justice respectively. Nor should this be surprising. If Rawls's argument and mine both simply limn the features of ideal choice, then we would expect nothing less than a measure of convergence in their conclusions.[1]

Department of Philosophy
1879 *Hall*
Princeton University
Princeton, NJ 08544-1006
USA
msmith@princeton

References

Nagel, Thomas 1970: *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

Parfit, Derek 1984: *Reasons and Persons*. Oxford: Clarendon Press.

Pettit, Philip, and Michael Smith 1997: 'Parfit's P'. In Jonathan Dancy (ed.), *Reading Parfit*, pp. 71–95. Oxford: Blackwell.

[1] A distant ancestor of this paper was given at Humean Ethics, The Second Warren Quinn Memorial Conference, UCLA, in 1998. I put the paper to one side after that, only returning to it in 2010–11 during a perfect sabbatical year spent in Berlin visiting the Humboldt University on a Humboldt Research Award courtesy of the Alexander von Humboldt Foundation. More recent versions were given as the Inaugural LEAP Lecture, Universitat Pompeu Fabra, Barcelona, in 2011; as a Keynote Address at Reasons of Love, Leuven, 2011; at the Department of Philosophy, Cambridge University, in 2011; at the Department of Philosophy Colloquium, Göttingen, in 2011; at Daniel Star's Boston University Ethics Reading Group in 2011; at Thomas Schmidt's Colloquium, Humboldt University, in 2011; at Tag der Handlungsgründe, Saarland University, in 2011; as the Claire Miller Lecture at the Chapel Hill Colloquium in 2011; at a meeting of Paper Tigers, Princeton University, in 2011; at the Summer Workshop on Rationalism at Melbourne University in 2012; at the Department of Philosophy Colloquium, University of California at Davis, in 2012; at the Department of Philosophy Colloquium, Fordham University, in 2012; at the Murphy Institute, Tulane University, in 2012; as one of three Hourani Lectures given at the University at Buffalo in 2012; at Practical Reason and Metaethics, a conference held at the University of Nebraska at Lincoln in 2012; and at a graduate seminar on reasons and rationality I gave at Princeton University in 2012. I received excellent comments on these occasions, so thanks to all of those who participated in the discussions. Special thanks are due to Joseph Biehl, David Braun, David Copp, Chris Cowie, Simon Cullen, Christoph Fehige, Elizabeth Harman, Barbara Herman, Richard Holton, Karen Jones, Katie Keene, Errol Lord, Barry Maguire, Julia Markovits, Thomas Mautner, Sven Nyholm, Joseph Raz, Geoff Sayre-McCord, Thomas Schmidt, Noel Swanson, Jay Wallace, Andrew Williams, and Daniel Wodak.

Proceedings of the Aristotelian Society, Vol. CXII*, Part 3*
doi: 10.1111/j.1467-9264.2012.00337.x

Rawls, John 1971: *A Theory of Justice*. Cambridge, MA: Harvard University Press.

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

——2011: 'Deontological Moral Obligations and Non-Welfarist Agent-Relative Values'. *Ratio*, 24, pp. 351–63.

Williams, Bernard 1981: 'Internal and External Reasons'. In his *Moral Luck*. Cambridge: Cambridge University Press.

——1995: 'Internal Reasons and the Obscurity of Blame'. Reprinted in his *Making Sense of Humanity*. Cambridge. Cambridge University Press.

Proceedings of the Aristotelian Society, Vol. CXII, *Part 3*
doi: 10.1111/j.1467-9264.2012.00337.x