Deontology and Descartes' Demon

Brian Weatherson, Cornell University

January 2007

1. Digesting Evidence

In his *Principles of Philosophy*, Descartes says,

> Finally, it is so evident that we are possessed of a free will that can give or withhold its assent, that this may be counted as on of the first and most ordinary notions that are found innately in us. (Descartes 1644/1965: xxix)

In this paper, I'm going to defend a broadly Cartesian position about doxastic freedom. At least some of our beliefs are freely formed, so we are responsible for them. Moreover, this has consequences for epistemology. But the *some* here is crucial. Some of our beliefs are not freely formed, and we are not responsible for those. And that has epistemological consequences too. Out of these considerations a concept of doxastic responsibility arises that is useful to the *externalist* in responding to several challenges. I will say at some length how it helps with the New Evil Demon problem, and I'll note some difficulties in reconciling internalism with the idea that justification is a kind of blamelessness. The internalist, I'll argue, has to say that justification is a kind of praiseworthiness, and this idea that praise is more relevant to epistemic concepts than blame will be a recurring theme of the paper.

While the kind of position I'm adopting has been gaining supporters in recent years, it is still largely unpopular. The arguments of William Alston (1988) have convinced many that it is a mistake to talk of doxastic freedom, or doxastic responsibility. The short version of this argument is that our beliefs are involuntary, and freedom and responsibility require voluntariness. The longer, and more careful, argument involves drawing some distinctions between ways in which we might come to be in a state. It helps to start with an example where the normative facts are relatively uncontroversial, namely digestion.

Imagine that Emma eats a meat pie, and due to a malfunction in her stomach the pie is not properly digested, leading to some medical complications. Is Emma responsible for her ill-health? Well, that depends on the back-story. If Emma knew that she could not properly digest meat pies, but ate one anyway, she is responsible for the illness via her responsibility for eating the pie. Even if Emma did not know this, she might be responsible for the state of her stomach. If her stomach could not digest the pie because it had been damaged by Emma's dietary habits, and let's say Emma knew that her diet could damage her stomach, then Emma is responsible for the state of her stomach and hence for the misdigestion of the pie and hence for her ill-health. But if neither of these conditions obtain, if it just happens that her stomach misdigests the pie, then Emma is not responsible for her ill-health. Even though the cause of her ill-health is something that *her* stomach does, he is not responsible for that since her stomach is not under her voluntary control. Put another way, her responsibility for maintaining her own health means that she is responsible for the type of digester she is, but he is not responsible for this token digestion.

Simplifying a little, Alston thinks that the case of belief is similar. Say that Emma has a false belief that *p*. Is she responsible for this piece of doxastic ill-health? Again, that depends on the back story. If Emma believes that *p* because she was careless in gathering evidence, and the evidence would have pointed to ~*p*, then she is responsible for being a bad gatherer of evidence. If Emma has been negligent in maintaining her doxastic health, or worse if she has been doing things she knows endangers doxastic health, then she is responsible for being the type of believer she is. But she is never responsible merely for the token belief that is formed. Her mind simply digests the evidence she has, and Emma's responsibility only extends to her duty to gather evidence for it, and her duty to keep her mind in good working order. She is not responsible for particular acts of evidential digestion.

But these particular acts of evidential digestion are the primary subject matters of epistemology. When we say Emma's belief is justified or unjustified, we usually mean that it is a good or bad response to the evidence in the circumstances. (I'm obviously here glossing over enormous disputes about what makes for a good response, what is evidence, and what relevance the circumstances have. But most theories of justification can be fit into this broad schema, provided we are liberal enough in interpreting the terms 'good', 'evidence' and 'circumstances'.) If Emma is not responsible for her response to the evidence, then either we have to divorce justification from responsibility, or we have to say that the concept of justification is defective.

We can summarise these considerations as a short argument. The following formulation is from Sharon Ryan (2003: 48).

> 1. If we have any epistemic obligations, then doxastic attitudes must sometimes be under our voluntary control.
> 2. Doxastic attitudes are never under our voluntarily control.
>
> ---
>
> 3. We do not have any epistemic obligations.

Ryan goes on to reject both premises. Matthias Steup (2000, 2005) also rejects both premises of this argument. I'm more sympathetic to premise 1, but I (tentatively) agree with them, against what sometimes seems to be orthodoxy, that premise 2 fails. That is, I endorse a kind of doxastic voluntarism. (Just what kind will become clearer as we go along.) There are four questions that anyone who endorses voluntarism, and wants to argue that this matters epistemologically, should I think answer. These are:

(A) What is wrong with current arguments against voluntarism?

(B) What does the voluntariness of (some) beliefs consist in?

(C) Which kinds of beliefs are voluntary?

(D) What difference does the distinction between these classes make for epistemology?

My answer to (A) will be similar to Ryan's, and to Steup's, but with I think enough differences in emphasis to be worth working through. My answer to (B), however, will be a little more different. I'm going to draw on some work on self-control to argue that some beliefs are voluntary because they are the result of exercises of, or failures to exercise, self-control. My answer to (C) is that what I'll call *inferential* beliefs are voluntary, while *perceptual* beliefs are not. Ryan and Steup sometimes seem to suggest that even perceptual beliefs are voluntary, and I don't think this is true. The consequence for this, I'll argue in answering (D), is that inferential beliefs should be judged by how well they respond to the evidence, while perceptual beliefs should be judged by how well they reflect reality.

I'll offer my answers to those four questions in sections 2, 4, 6 and 7. In between I'll discuss recent work on self-control (section 3) and the contrast between my answer to (B) and other voluntarist answers (section 5). In section 8 I'll say how my partially voluntarist position gives the externalist a way to avoid the New Evil Demon problem. And in section 9 I'll make a direct argument for the idea that justification is a kind of praiseworthiness, not a kind of blamelessness.

Before we start, I want to note two ways, other than Ryan's, of formulating an argument against doxastic responsibility. These are going to seem quite similar to Ryan's formulation, but I think they hide important differences. The first version uses the idea that some doings (or states) are volitional. That is, we do them (or are in them) because we formed a volition to do so, and this volition causes the doing (or state) in the right kind of way.

1. If we have any epistemic obligations, then either the formation or maintenance of doxastic attitudes must sometimes be volitional.
2. The formation or maintenance of doxastic attitudes is never volitional.

---

3. We do not have any epistemic obligations.

I won't argue against premise 2 of this argument, though Carl Ginet (1985, 2001) has done so. But I think there's little to be said for premise 1. The principle behind it is that we are only responsible for volitional doings. And that principle is very dubious. We could run the kind of regress arguments against it that Gilbert Ryle (1949) offers. But it is simpler to note some everyday counterexamples. Borrowing an example from Angela M Smith (2005), if I forget a friend's birthday, that is something I am responsible and blameworthy for, but forgetting a birthday is not volitional. (Below I'll offer a Rylean argument that we are sometimes praiseworthy for doings that are not volitional.) So this argument fails. Alternatively, we could run the argument by appeal to freedom.

> 1. If we have any epistemic obligations, then doxastic attitudes must sometimes be free.
> 2. Doxastic attitudes are never free.
>
> ---
>
> 3. We do not have any epistemic obligations.

Premise 1 of this argument is more plausible. But, as we'll see presently, premise 2 is not very plausible. Whether Descartes was right that premise 2 is obviously false, it does seem on reflection very hard to defend. So this argument fails. Ryan's formulation is interesting because it isn't clear just which of the premises fails. As I said, I'm going to suggest that premise 2 fails, and that doxastic attitudes are voluntary. But this will turn on some fine judgments about the voluntary/involuntary boundary. If I'm wrong about those judgments, then the arguments below will suggest that premise 1, not premise 2, in Ryan's formulation fails. Either way though, the argument is unsuccessful.

2. Responding to the Involuntarists

There are two kinds of argument against the idea that belief is voluntary. One kind, tracing back to Bernard Williams (1973), holds that the possibility of voluntary belief can be shown to be incoherent by reflection on the concept of belief. This argument is no longer widely endorsed. Nishi Shah (2002) provides an excellent discussion of the problems with Williams' argument. I will focus on the other kind, that claims we can see that belief is involuntary by observing differences between beliefs and paradigm cases of voluntary actions. I will make three objections to these arguments. First, the argument looks much less plausible once we distinguish between having a belief and forming a belief. Second, the argument seems to rely on inferring from the fact that we don't do something (in particular, believe something that we have excellent evidence is false) to the conclusion that we can't do it. As Sharon Ryan (2003) points out, it might simply be that we won't do it. Third, the argument relies on too narrow a conception of what is voluntary, and when we get a more accurate grasp on that concept, we'll give up the argument. Here is a representative version of the argument from William Alston.

> Can you, at this moment, start to believe that the United States is still a colony of Great Britain, just by deciding to do so? … [S]uppose that someone offers you $500,000,000 to believe it, and you are much more interested in the money than in believing the truth. Could you do what it takes to get that reward? . . . Can you switch propositional attitudes toward that proposition just by deciding to do so? It

> seems clear to me that I have no such power. Volitions, decisions, or choosings don't hook up with anything in the way of propositional attitude inauguration, just as they don't hook up with the secretion of gastric juices or cell metabolism. (Alston 1989: 263)

Now Alston does note, just one page earlier, that what is really relevant is whether our being in a state of belief is voluntary, not whether the activity of belief formation is voluntary. But he thinks nevertheless that issues about whether we can form beliefs, any old beliefs it seems, voluntarily matters to the question about the voluntariness of belief states.

If we think about what it is to be in a state voluntarily, this all seems beside the point. We can see this by considering what it is to be in a political state voluntarily. Consider Shane, who was born into Victoria. His coming to be in Victoria was hence not, in any way, voluntary. Shane is now a grown man, and he has heard many travellers' tales of far away lands. But the apparent attractions of Sydney and other places have no pull on Shane; he has decided to stay in Victoria. If he has the capacity to leave Victoria, then Shane's continued presence in Victoria is voluntary. Similarly, we are voluntarily in a belief state if we have the capacity to leave it, but choose not to exercise this capacity. Whether the belief was formed voluntarily is beside the point.

If Shane leaves a state, the natural place to leave is for another state, perhaps New South Wales or South Australia. It might be thought that if we leave a belief state, we have to move into another belief state. So to have this capacity to leave, we need the ability to form beliefs voluntarily. Not at all. The capacity to become uncertain is capacity enough. (If Shane has a boat, and the capacity to flourish at sea, then perhaps he too can have the capacity to leave Victoria without the capacity to go into another state.)

But do we have the capacity to become uncertain? Descartes thought so; the point of the First Meditation is to show us how to exercise this capacity. Moreover, this capacity need not be one that we exercise in any particularly nearby possible worlds. We might exercise our freedom by always doing the right thing. As Descartes goes on to say in the Fourth Meditation.

> For in order to be free, there is no need for me to be capable of going in each of two directions; on the contrary, the more I incline in one direction – either because I clearly understand that reasons of truth and goodness point that way, or because of a divinely produced disposition of my inmost thoughts – the freer is my choice. (Descartes 1641/1996: 40)

This seems like an important truth. A roulette wheel is not free, and nor is where it ends up a matter of its own voluntary control. If we want to be free, we should not aim to resemble a roulette wheel. John Cottingham (2002) interprets Descartes here as suggesting that there is a gap between free action and voluntary action, contrasting his "strongly compatibilist notion of human freedom" (350) with the "doxastic involuntarism" (355) suggested by the following lines of the Third Meditation.

> Yet when I turn to the things themselves which I think I perceive very clearly, I am so convinced by them that I spontaneously declare: let whoever can do so deceive me, he will never bring it about that I am nothing, so long as I continue to think that I am something … (Descartes 1641/1996: 25)

Now there are two questions here. The first is whether Descartes intended to draw this distinction. That is, whether Descartes thought that the kind of free actions that he discusses in the Fourth Meditations, the free action where we are incapable of going in the other directions, are nevertheless involuntary. I don't have any informed opinions about this question. The second is whether this kind of consideration supports the distinction between the free and the voluntary. And it seems to me that it does not. Just as Descartes says the free person will be moved by reasons in the right way, it seems natural to say that a person who acts voluntarily will be responsive to reasons. Voluntary action does require freedom from certain kinds of coercion, but the world does not coerce us when it gives us reason to believe one thing rather than another. If we have voluntary control over our beliefs, then we should be compelled by the sight of rain to believe it is raining.

In her discussion of the puzzle of imaginative resistance, Tamar Szabó Gendler (2000) notes that philosophers have a tendency to read too much into intuitions about certain cases. What we can tell from various thought experiments is that in certain circumstances we won't do a certain thing. But getting from what we won't do to what we can't do is a tricky matter, and it is a bad mistake to infer from won't to can't too quickly. Matthias Steup (2000) points out that if we try to stick a knife into our hand, we similarly won't do it. But this is no evidence that we cannot do it. And Sharon Ryan (2003) notes that we will not bring ourselves to run over pedestrians for no reason. For most of us, our moral sense prevents acting quite this destructively. Yet our continued avoiding of pedestrians is a series of free, even voluntary, actions. We could run over the pedestrians, but we will not. Since forming false beliefs is a form of self-harm, it is not surprising that it has a similar phenomenology, even if it is genuinely possible.

It might be argued that we will engage in small forms of self-harm that we can do when the financial rewards are great enough. So we should be able to form this belief about the United States for a large amount sum of money. But I suspect that the only way to exercise the capacity to believe the United States is still a colony is by first suspending my belief that it is no longer a colony. And the only way I can do that is by generally becoming more sceptical of what I've been told over the years. Once I get into such a sceptical mood, I will be sceptical of claims that I'll get half a billion dollars should I have this wild political belief. So I won't form the belief in part because the 'promisor' lacks the capacity to sufficiently convince me that I'll be richly rewarded for doing so. This looks like a lack of capacity on their part, not my part.

The final point to make about this argument, and those like it, is that if we are to conclude that belief formation is *never* voluntary, then we need to compare it to all kinds of voluntary action. And Alston really only ever compares belief formation to volitional action. If this doesn't exhaust the range of voluntary action, then belief formation might be properly analogous to some other voluntary action. Indeed, this turns out to be the case. To see so, we need to make a small detour through modern work on self-control.

3. How to Control Your Temper

To start, let's consider three examples of a person failing to keep a commitment they've made about what the good life is. The three ways will be familiar from Gary Watson's discussion of recklessness, weakness and compulsion (Watson 1977), and the discussion of these cases by Jeanette Kennett and Michael Smith (1994, 1996). My characterisation of the cases will turn out to differ a little from theirs, but the cases are similar. Each of the examples concerns a character Murray, who has decided that he should not swear around his young son Red. He resolves to do this, and has been working on curbing his tendency to swear whenever anything bad happens. But three times over the course of the day he breaks his commitment.[1]

The first time comes when Murray puts his hand down on a hot plate that he did not realise was on. The searing pain undermines his self-control, and he is unable to stop himself from swearing loudly through the pain.

The second time comes when Murray drops and breaks a wine glass. Murray doesn't lose his self-control, but he doesn't exercise the self-control he has. He temporarily forgets his commitment and so, quite literally, curses his misfortune. On doing so he immediately remembers that Red is around, and the commitment he has made, and regrets what he did.

The third time comes on the tram home, when Murray gets into a disagreement with a political opponent. Murray can't find the words to express what he feels about the opponent without breaking his commitment. So he decides, without much reason, that his need to express what he feels outweighs his commitment, and starts describing his opponent using language he would, all things considered, not have used around young Red.

The first and third cases are close to textbook cases of compulsion and recklessness. Note in the first case that when Murray reflects back on what happened, he might be irritated that his work on reducing his tendency to swear has not been more successful. But he will not be upset that he didn't exercise more self-control on that occasion. He didn't have, no normal person would have, the amount of self-control he would have needed to stop swearing then. All that would help is having the disposition to say different things when his self-control is defeated. And that's not a disposition he can acquire on the spot.

---

[1] The cases, especially the second, were inspired by Richard Holton's (2003, 2004) discussion of resolutions to prevent 'automatic' actions like smoking or sleeping in.

I've described the first case as one where Murray's self-control is undermined. This is a term taken from recent work by Richard Holton and Stephen Shute (forthcoming), who carefully distinguish between self-control being *undermined* by a provocation, and it being *overwhelmed* by a provocation. Undermining occurs when the provocation causes the agent to have less self-control than they usually have; overwhelming occurs when the provocation is too much for the agent's control. The difference is relevant to them, because they are interested in what it is for an agent to *lose* control. That seems to be what happens here. After all, the things one would naturally do afterwards (jumping around, screaming, swearing if one's so disposed) don't seem particularly controlled by any measure.

Similarly I've accepted Watson's desrciption of cases like the third as instances of recklessness, but we shouldn't think this necessarily contrasts with weakness. It might be that in this case Murray is both weak and reckless. He is not akratic, if we stipulatively define akrasia as acting against one's better judgment. But if we accept Richard Holton's view that weakness of will consists in being "too ready to reconsider their intentions" (page 241), then in this case Murray is weak-willed.[2] This seems to be the right way to talk about the case to me. With these details in place, we can talk about what's crucial to this essay, the contrast with the second case.

In the second case Murray fails to exercise self-control. He could have prevented himself from swearing in front of his son. Breaking a wine glass is irritating, but it neither undermines nor, necessarily, overwhelms self-control. Murray had the capacity to think about his resolution to not swear in front of Red. And if he had exercised this capacity, he would not have sworn when he did.

In the first case, Murray will only regret his lack of prior work at changing his dispositions in cases where his control fails. In the second case he will regret that, but he will also regret what he did on that occasion, for he could have kept his resolution, had only he thought of it. This regret seems appropriate, for in the second case he did something wrong at the time he swore, as well perhaps as having done something wrong earlier. (Namely, not having worked hard enough on his dispositions.) This difference in regret doesn't constitute the difference between compulsion and a case where self-control fails, but it is pretty good evidence that this is a failure of self-control.

So the second case is not one where Murray was compelled. He had the capacity to keep his commitment, and nothing was stopping him exercising this control, but he failed to do so. His failure was a failure of self-control. Murray's self-control is, in this case, overwhelmed by the provocation. But it need not have been. Within some fairly broad limits, how much self-control we exercise is up to us.[3] Murray's failure of self-control is culpable because anyone with the capacity for self-control Murray has could have avoided breaking his commitment. I'm not going to try to offer an analysis of what it is to have a capacity, but I suspect something like the complicated counterfactual analysis Kennett and Smith offer, and that Smith offers elsewhere (Smith 1997, 2003), is broadly correct.[4]

---

[2] Whether Murray is akratic is a slightly more complicated question than I've suggested in the text. If akrasia is acting against one's judgment, then he is not; if akrasia is acting against one's *considered* judgment, then he is. 'Akrasia' is a technical term, so I don't think a huge amount turns on what we say about this question.

There is an interesting historical precedent for Holton's theory of weakness of will. Ryle hints at a similar position to Holton's when he says "Strength of will is a propensity the exercise of which consist in sticking to tasks' that is, in not being deterred or diverted. Weakness of will is having too little of this propensity." (1949: 73) But the idea is not well developed in Ryle. We'll return below to the differences between Ryle's and Holton's theories.

Kennett and Smith stress two things about this capacity that are worth noting here. First, having this kind of capacity is part of what it is to be rational. That is, being rational requires thinking of the right thing at the right time. As Ryle says, "Intelligently reflecting how to act is, among other things, considering what is pertinent and disregarding what is inappropriate." (1949: 31) Second, Kennett and Smith note that exercises of this capacity cannot be volitional. Following Davidson (1963), they say they cannot be *actions*. I find this terminology somewhat strained. Catching a fast moving ball is an action, I'd say, but it doesn't seem to be volitional. So I'll use 'volitional action' for this Davidsonian sense of action.

Many recent philosophers have endorsed the idea that some of the mental states for which we hold people responsible are not voluntary, or at least are not volitional. Adams (1985), Heller (2000), Owens (2000: Ch. 8) and Hieronymi (forthcoming) note ways in which we appropriately blame people for being in certain states, where being in that state is not volitional. Something like this idea seems to be behind Ryle's several regress arguments against the intellectualist legend. It just isn't true that what we do divides cleanly into outcomes of conscious thought on the one hand, and mere bodily movements (a la digestion) on the other.[5] Rather there is a spectrum of cases from pure ratiocination at one end to pure bodily movement at the other. And some of the things in the middle of this spectrum are proper subjects of reactive attitudes. The focus in this literature has been on blame, but some states in the middle of this spectrum are also praiseworthy.

Consider some action that is strikingly imaginative, e.g. a writer's apt metaphor or, say, a cricket captain's imaginative field placements. It seems that, assuming the field settings are successful, the captain deserves praise for being so imaginative. But of course the captain did not, really could not, first intend to imagine such field settings, then carry out that intention. So something for which the captain deserves praise, his act of imagination, is not volitional. So not all praiseworthy things we do are volitional.

There are two responses to this argument that I can imagine, neither of them particularly plausible. First, we might think that the captain's imagination is simply a remarkable feature of nature, as the Great Barrier Reef is. It is God, or Mother Nature, who should be praised, not the captain. Now it seems fair to react to some attributes of a person this way. A person doesn't deserve praise for having great eyesight, for example. But such a reaction seems grossly inappropriate, almost dehumanising, in this case. To be sure, we might also praise God or Mother Nature for yielding such an imaginative person, but we'll do that *as well as* rather than *instead of*, praising the person. Second, we might praise the captain for his work in studying the game, and thinking about possible ways to dismiss batsmen, rather than this particular action. But if that's what we praise the captain for, we should equally praise the captain's opponent, a hard working dullard. And that doesn't seem right. The hard-working dullard deserves praise for his hard work in the lead up, but the hard-working imaginative skipper deserves praise for what he does in the game too. So reactive attitudes, particularly praise, are appropriately directed at things people do even if these things aren't volitional.

---

[3] Holton (2003) compares self-control to a muscle that we can exercise. We can make a similar point to the one in the text about physical muscles. If I try to lift a box of books and fail, that doesn't show I lack the muscular capacity to lift the box; I might not have been trying hard enough.
[4] Ryle (1949: 71ff) also offers a counterfactual account of capacities that seems largely accurate.
[5] As I read him, Ryle takes this fact to reveal an important weakness in Descartes' theory of mind.

The key point of this section then is that responsibility outruns volition. Some actions are blameworthy because they are failures of self-control. Some actions are praiseworthy because they are wonderful feats of imagination. But neither failing to exercise self-control, nor exercising imagination, needs be volitional is order to be a locus of responsibility. I'll argue in the next section that these considerations support the idea of responsibility for beliefs.

4. Voluntariness about Belief

Here's a situation that will seem familiar to anyone who has spent time in a student household. Mark is writing out the shopping list for the weekly grocery shop. He goes to the fridge and sees that there is a carton of orange juice in the fridge. He forms the belief that there is orange juice in the fridge, and hence that he doesn't need to buy orange juice. As it turns out both of these beliefs are false. One of his housemates finishes off the orange juice, but stupidly put the empty carton back in the fridge. When Mark finds this out, he is irritated at his housemate, but he is also irritated at himself. He didn't have to draw the conclusion that there was orange juice in the fridge. He was, after all, living in a student house where people do all sorts of dumb things. That his housemate might have returned an empty container to the fridge was well within the range of live possibilities. Indeed had he even considered the possibility he would have thought it was a live possibility, and checked whether the container was empty before forming beliefs about what was needed for the shopping.

Examples like this can be easily multiplied. There are all sorts of beliefs that we form in haste, where we could have stopped to consider the various realistic hypotheses consistent with the evidence, and doing so would have stopped us forming the belief. Indeed, unless one is a real master of belief formation, it shouldn't be too hard to remember such episodes frequently from one's everyday life. These conclusions that we leap to are voluntary beliefs; we could have avoided forming them. And this isn't the only way that voluntary doings, like calling a relevant possibility to mind, can matter to belief. The next example will be a little more controversial, but it points at the importance of dismissing irrelevant possibilities.

Later that evening, Mark is watching his team, Geelong, lose another football game. Geelong are down by eight goals with fifteen minutes to go. His housemates are leaving to go see a movie, and want to know if Mark wants to come along. He says that he's watching the end of the game because Geelong might come back. One of his housemates replies, "I guess it's possible they'll win. Like it's possible they'll call you up next week to see if you want a game with them." Mark replies, "Yeah, you're right. This one's over. What's the movie?" Mark doesn't just give in to his housemates, he forms the belief that Geelong will lose. Later that night, when asked what the result of the game was, he says that he didn't see the final score, but that Geelong lost by a fair bit. (In Weatherson 2005 I go into a lot more detail on the relation between not taking possibilities seriously, and having beliefs. The upshot is that what Mark does can count as belief formation, even if his credence that Geelong will lose does not rise.)

Now it is tempting, or perhaps I should say that *I* am tempted, to view the housemate as offering Mark a reason to believe that Geelong will lose. We could view the housemate's comments as shorthand for the argument that Geelong's winning is as likely as Mark's playing for Geelong, and since the latter won't happen, neither will the former. And maybe that's *part* of what the housemate is doing. But the larger part is that she's *mocking* Mark for his misplaced confidence. And the point of mocking someone, at least the point of constructive mockery like this, is to get them to change their attitudes. Mark does so, by ceasing to take seriously the possibility that Geelong will come back. In doing so, he exercises a capacity he had for a while, the capacity to cease taking this unserious possibility seriously, but needed to be prompted to use.

In both cases I say Mark's belief formation is voluntary. In the first case he forms the belief because he does not exercise his doxastic self-control. He should have hesitated and not formed a belief until he checked the orange juice. And he would have done so if only he'd thought of the possibility that the container was empty. But he didn't. And just as things we do because we don't bring the right thing to mind, like Murray's swearing in the second case, are voluntary and blameworthy, Mark's belief is voluntary and blameworthy. In the second case, he forms the belief by ceasing to take an unserious possibility seriously. In most cases of non-perceptual, non-testimonial belief formation, there is a counter-possibility that we could have taken seriously. Skill at being a believer involves not taking extreme possibilities, from Cartesian sceptical scenarios to unlikely footballing heroics, seriously. Exercises of such skill are rarely, if ever, volitional. But just like other mental activities that aren't volitional can be voluntary and praiseworthy, not taking an extreme possibility seriously can be voluntary and praiseworthy.[6]

---

[6] Ryle (1949: 29ff) stresses the importance of calling the right things to mind to rational thought and action. I'm using a case here where Mark deliberately casts an option from his mind, but the more general point is that what possibilities we call to mind is a crucial part of rational action, and can be praiseworthy or blameworthy, whether or not it is volitional.

I've made two claims for Mark's beliefs in the above two cases. First, they are instances of voluntary belief formation. In each case he could have done otherwise, either by exercising or failing to exercise his capacity to take various hypotheses seriously. Second, they are appropriate subjects of praise and blame. I imagine some people will agree with the second point but not the first. They will say that only volitional actions are voluntary, even though things we do like bringing relevant considerations to mind are praiseworthy or blameworthy. Such people will agree with most of what I say in this paper. In particularly they'll agree that the examples involving Mark undermine Alston's argument against the applicability of deontological concepts in epistemology. So I'm not going to die in a ditch over just what we call voluntary. That is, I won't fuss too much over whether we want to say premise 2 in Ryan's formulation of the argument is shown to be false by these examples (as I say) or premise 1 is shown to be false (as such an objector will say.) I will just note that it is hard for such people to say intuitive things about the second instance of Murray's swearing, and this seems like a strong reason to not adopt their position.[7]

5. Ryan and Steup

Sharon Ryan has a slightly different view. She thinks that the truth of voluntarism consists in the fact that we hold certain beliefs intentionally. She doesn't offer an analysis of what it is to do something intentionally, except to say that consciously deciding to do something is not necessary for doing it intentionally, but doing it purposefully is. (2003: 70-1). In a similar vein, she says "When there's a car zooming toward me and I believe that there is, I'm believing freely because I'm believing what I mean to believe." (74) This is said to be an intentional, and I take it a voluntary, belief.

It seems to me that there's a large difference between things we voluntarily do, and things we mean to do, or do purposefully. There are several things we do voluntarily without meaning to do them. Murray's swearing in the second example above is one instance. When we misspeak, or (as I frequently do) mistype, we do things voluntarily without meaning to do them. I don't mean by mistype cases where we simply hit the wrong key, but such cases as where I write in one more negation than I meant to, or, as I did earlier this evening, write "S is justified in believing that *p*" when I meant to write "S is justified in believing that she's justified in believing that *p*." These are voluntary actions because I had the capacity to get it right, but didn't exercise the capacity. But they aren't things I meant to do. (I suspect there are also cases where we do things because we mean to do them, but they aren't voluntary. These include cases where we train ourselves to produce a reflexive response. But I won't stress such cases here.)

---

[7] Ryle seems to have taken an intermediate position. He holds, I think, the view that voluntary acts are culpable acts where we had the capacity to do otherwise (71). So Mark's belief about the orange juice is voluntary because he had the capacity to retain doubt, and nothing prevented him exercising it. But the belief about the football is not voluntary because we shouldn't talk about praiseworthy acts being voluntary or involuntary. The last point is the kind of error that Grice (1989: Ch 1) showed us how to avoid.

Matthias Steup (forthcoming) argues that if compatibilism is true about free action, then our beliefs are free. His argument consists in running through the most plausible candidates to be compatibilist notions of freedom and showing that each is either mistaken or implies that beliefs are free. I agree with a lot of what Steup says, indeed this paper has been heavily influenced by what he says. But one crucial analogy fails I think. Steup is concerned to reject the premise that if Φ-ing is free, one Φs because one has formed the intention to Φ. His response centres around 'automatic' actions, such as the things we do when starting our drive to work: inserting the key, shifting into reverse, etc.

> The question is whether they are caused by any antecedently formed intentions. I don't think they are. … I didn't form an intention to … shift into reverse…. I do things like that automatically, without thinking about them, and I assume you do too. But one can't form an intention to Φ without thinking about Φing … Just one more example: I'd like to see the person who, just before brushing her teeth, forms the intention to unscrew the cap of the toothpaste tube. (11-12)

I suspect that Steup simply has to look in the mirror. It's true that we don't usually form *conscious* intentions to shift into reverse, or unscrew the cap, but not all intentions are conscious. If we were asked later, perhaps by someone who thought we'd acted wrongly, whether we intended to do these things, the natural answer is *yes*. The best explanation of this is that we really did have an intention to do them, albeit an unconscious one. (I'm indebted here to Ishani Maitra.)

Steup is right that free actions don't require a prior intention, but his examples don't quite work. The examples I've used above are the Rylean regress stoppers, such as acts of imagination, and actions that we do *because* we didn't think, like Murray's swearing. If asked later whether he intended to say what he said, Murray would say *yes* in the third example, but (I think) *no* in the first and second. Intuitively, I think, he didn't have such an intention.[8]

6. Involuntarism about Perceptual Beliefs

In some early 1990s papers, Daniel Gilbert and colleagues defended a rather startling thesis concerning the relation of comprehension and belief. Casual introspection suggests that when one reads or hears something, one first comprehends it and then, if it is backed by sufficient reasons, believes it. Gilbert (1991) argues against this seeming separation of comprehension and belief, and in favour of a view said to derive from Spinoza. When we comprehend a sentence, we add it to our stock of beliefs. If the new belief is implausible given our old beliefs, then we "unbelieve" it. (The experimental evidence for this is set out in greater detail in Gilbert et al 1990 and Gilbert et al 1993.)

---

[8] If so, Murray is not weak-willed according to Holton's theory of will, but, since he doesn't keep his resolution, he is weak-willed according to Ryle's otherwise similar theory. This seems to be an advantage of Holton's theory over Ryle's. Murray's problem isn't that his will was weak, it's that it wasn't called on. More generally, Ryle's identification of weakness of will with irresoluteness seems to fail for people who frequently *forget* their resolutions. These people are surely irresolute, but (in agreement with Holton's theory) I think they aren't weak-willed.

We may picturesquely compare the two models of belief and comprehension to two models for security. The way security works at a nightclub is that anyone can turn up at the door, but only those cleared by the guards are allowed in. On the other hand, the way security works at a shopping mall is that anyone is allowed in, but security might remove those it regards as undesirable. Intuitively, our minds work on the nightclub model. A hypothesis can turn up and ask for admission, but it has to be approved by our cognitive security before we adopt it as a belief. Gilbert's position is that we work on the shopping mall model. Any hypothesis put in front of us is allowed in, as a belief, and the role of security is to remove troublemakers once they have been brought inside.

Now I don't want to say Gilbert's theory is correct. (The experimental evidence for it is challenged in Hasson et al 2006.) But I do want to argue that if it is correct, then there is a kind of belief that is clearly involuntary. We don't have much control over what claims pass in front of our eyes, or to our ears. (We have some indirect control over this – we could wear eye shades and ear plugs – but no direct control, which is what's relevant.) If all such claims are believed, these are involuntary beliefs. To be sure, nothing Gilbert says implies that we can't quickly regain voluntary control over our beliefs as we unbelieve the unwanted inputs. But in the time it takes to do this, our beliefs are out of our control.

Gilbert's theory is rather contentious, but there are other kinds of mental representations that it seems clear we can't help forming. In *The Modularity of Mind*, Jerry Fodor has a long discussion of how the various input modules that he believes to exist are not under our voluntary control. (As he says, they have a mandatory operation. See pages 52 – 55 in particular, but the theme is central to the book.) If I'm sitting on a train opposite some people who are chatting away, I can't help but hear what they say. (Unless, perhaps, I put my fingers in my ear.) This is true not just in the sense that I can't help receive the sound waves generated by their vocalisations. I also can't help *interpret* and *comprehend* what they are saying. Much as I might like to not be bothered with the details of their lives, I can't help but hear what they say *as* a string of English sentences. Not just hearing, but hearing *as* happens automatically.

This automatic 'hearing as' is not under my voluntary control. I do not do it because I want to do it, or as part of a general plan that I endorse or have chosen to undertake. It does not reflect any deep features of my character. (Frankly I'd much rather that I just heard most of these conversations as meaningless noise, like the train's sound.) But I do it, involuntarily, nonetheless. This involuntariness is reflected in some of our practices. A friend tells me not to listen to X, because X is so often wrong about everything. Next I see the friend I say that I now believe that *p*, and when the friend asks why, I say it is because X said that *p*. The friend might *admonish* me.

They won't admonish me for being within hearing range of X; that might have been unavoidable. And, crucially, they won't admonish me for interpreting X's utterances. Taken literally, that might be what they were asking me not to do. But they'll know it was unavoidable. What they were really asking me not to do was the one relevant thing that I had control over, namely believe what X said.

As Fodor points out at length, both seeing as and hearing as are generally outside voluntary control. Our perceptual systems, and by this I'm including verbal processing systems, quickly produce representations that are outside voluntary control in any sense. If any of these representations amount to beliefs, then there are some involuntary beliefs that we have. So we might think that in the case above, although it was up to me to believe that *p*, it wasn't up to me to believe that, say, X said that *p*, because this belief was produced by a modular system over which I have no control.

This is not the position that Fodor takes. He thinks that *beliefs* are not produced by input modules. Rather, the non-modular part of the mind, the central processor, is solely responsible for forming and fixing beliefs. And the operation of this central processor is generally not mandatory, at least not in the sense that the operation of the modules is mandatory. Whether this is right seems to turn (in part) on a hard question to do with the analysis of belief.

Let's quickly review Fodor's views on the behaviour of input modules. The purpose of each module is to, within a specified domain, quickly and automatically produce representations of the world. These are, as on the nightclub model, then presented to cognition to be allowed in as beliefs or not. Here's how Fodor puts it.

> I am supposing that input systems offer central processes hypotheses about the world, such hypotheses being responsive to the current, local distribution of proximal stimulations. The evaluation of these hypotheses in light of the rest of what one knows is one of the things that central processes are for; indeed, it *is* the fixation of perceptual belief. (1983: 136)

But these representations don't just offer hypotheses. They can also guide action prior to being 'approved' by the central processes. That, at least, seems to be the point of Fodor's discussion of the evolutionary advantages of having fast modules (1983: 70-71). The core idea is that when one is at risk of being eaten by a panther, there is much to be said for a quick, automatic, panther recognition device. But there is just as much to be said for acting immediately on one's panther recognition capacities rather than, say, searching for possible reasons why this panther appearance might be deceptive. And browsing reason space for such evidence of deceptions is just what central processes, in Fodor's sense, do. So it seems the natural reaction to seeing a panther should be, and is, guided more-or-less directly by the input modules not central processes.

So these 'hypotheses' are representations with belief-like direction of fit, i.e. they are responsive to the world, that guide action in the way that beliefs do. These are starting to sound a lot like beliefs. Perhaps we should take a Gilbert-style line and say that we automatically believe what we perceive, and the role of Fodorian central processes is not to accept or reject mere hypotheses, but to unbelieve undesirable inputs.[9] There are a number of considerations that can be raised for and against this idea, and perhaps our concept of belief is not fine enough to settle the matter. But let's first look at three reasons for thinking these inputs are not beliefs.

First, if they are beliefs then we are often led into inconsistency. If we are looking at a scene we know to be illusory, then we might see something as an *F* when we know it is not an *F*. If the outputs of visual modules are beliefs, then we inconsistently believe both that it is and is not *F*. Perhaps this inconsistency is not troubling, however. After all, one of the two inconsistent beliefs is involuntary, so we are not responsible for it. So this inconsistency is not a sign of irrationality, just a sign of defective perception. And that's not something we should be surprised by; the case by definition is one where perception misfires.

Second, the inputs don't, qua inputs, interact with other beliefs in the right kind of way. Even if we believe that *if p then q*, and perceive that *p*, we won't even be disposed to infer that *q* unless and until *p* gets processed centrally. On this point, see Stich (1978) and Fodor (1983: 83-6). The above considerations in favour of treating inputs as beliefs turned heavily on the idea that they have the same functional characteristics as paradigm beliefs. But as David Braddon-Mitchell and Frank Jackson (2007: 114-23) stress, functionalism can only be saved from counterexamples if we include these inferential connections between belief states in the functional charactisation of belief. So from a functionalist point of view, the encapsulation of input states counts heavily against their being beliefs.

Finally, if Fodor is right, then the belief-like representation of the central processes form something like a natural kind. On the other hand, the class consisting of these representations plus the representations of the input modules looks much more like a disjunctive kind. Even if all members of the class play the characteristic role of beliefs, we might think it is central to our concept of belief that belief is a natural kind. So these inputs should not count as beliefs.

On the other hand, we shouldn't overestimate the role of central processes, even if Fodor is right that central processes are quite different to input systems. There are two related features of the way we process inputs that point towards counting some inputs as beliefs, and hence as involuntary beliefs. The first feature is that we don't have to put any effort into believing what we see. On the contrary, as both Descartes and Hume were well aware, we believe what we see by default, and have to put effort into being sceptical. The second feature is that, dramatic efforts aside, we can only be so sceptical. Perhaps sustained reflection on the possibility of an evil demon can make us doubt all of our perceptions at once. But in all probability, at least most of the time, we can't doubt everything we see and hear.[10] We can perhaps doubt any perceptual input we receive, but we can't doubt them all.

---

[9] To be clear, the position being considered here is not that we automatically believe *p* when someone says *p* to us, but that we automatically believe that they said that *p*.

In the picturesque terms from above, we might think our security system is less like a nightclub and more like the way customs appears to work at many airports. (Heathrow Airport is especially like this, but I think it is not that unusual.) Everyone gets a cursory glance from the customs officials, but most people walk through the customs hall without even being held up for an instant, and there aren't enough officials to stop everyone even if they wanted to. Our central processes, faced with the overwhelming stream of perceptual inputs, are less the all-powerful nightclub bouncer and more the overworked customs official, looking for the occasional smuggler who shouldn't be getting through.

The fact that inputs turn into fully fledged beliefs by default is some reason to say that they are beliefs as they stand. It's noteworthy that what Gilbert et al's experiments primarily tested was whether sentences presented to subjects under cognitive load ended up as beliefs of the subjects. Now this could be because comprehending a sentence implies, at least temporarily, believing it. But perhaps a more natural reading in the first instance is that inputted sentences turn into beliefs unless we do something about it. Gilbert et al are happy inferring that in this case, the inputs *are* beliefs until and unless we do that something. This seems to be evidence that the concept of belief philosophers and psychologists use include states that need to be actively rejected if they are not to acquire all the paradigm features of belief. And that includes the inputs from Fodorian modules.

That argument is fairly speculative, but we can make more of the fact that subjects can't stop everything coming through. This implies that there will be some long disjunctions of perceptual inputs that they will end up believing no matter how hard they try. Any given input can be rejected, but subjects only have so much capacity to block the flow of perceptual inputs. So some long disjunctions will turn up in their beliefs no matter how hard they try to keep them out. I think these are involuntary beliefs.

So I conclude tentatively that perceptual inputs are involuntary beliefs, at least for the time it would take the central processes to evaluate them were it disposed to do so. And I conclude less tentatively that subjects involuntarily believe long disjunctions of perceptual inputs. So some beliefs are involuntary.

Space considerations prevent a full investigation of this, but there is an interesting connection here to some late medieval ideas about evidence. In a discussion of how Descartes differed from his medieval influences, Matthew L. Jones writes "For Descartes, the realignment of one's life came about by training oneself to assent only to the evident; for the scholastics, assenting to the evident required no exercise, as it was automatic." (2006: 84)[11] There is much contemporary interest in the analysis of evidence, with Timothy Williamson's proposal that our evidence is all of our knowledge being a central focus (Williamson 2000: Ch. 9). I think there's much to be said for using Fodor's work on automatic input systems to revive the medieval idea that the evident is that which we believe automatically, or perhaps it is those pieces of knowledge that we came to believe automatically. As I said though, space prevents a full investigation of these interesting issues.

---

[10] As noted in the last footnote, when I talk here about what we hear, I mean to include propositions of the form *S said that p*, not necessarily the *p* that *S* says.

7. Epistemological Consequences

So some of our beliefs, loosely speaking the perceptual beliefs, are spontaneous and involuntary, while other beliefs, the inferential beliefs, are voluntary in that we have the capacity to check them by paying greater heed to counter-possibilities. (In what follows it won't matter much whether we take the spontaneous beliefs to include all the perceptual inputs, or just the long disjunctions of perceptual inputs that are beyond our capacity to reject. I'll note the few points where it matters significantly.) This has some epistemological consequences, for the appropriate standards for spontaneous, involuntary beliefs are different to the appropriate standards for considered, reflective beliefs. I include in the latter category beliefs that were formed when considered reflection was possible, but was not undertaken.

To think about the standards for spontaneous beliefs, start by considering the criteria we could use to say that one kind of animal has a better visual system than another. One dimension along which we could compare the two animals concerns discriminatory capacity – can one animal distinguish between two things that the other cannot distinguish? But we would also distinguish between two animals with equally fine-grained visual representations, and the way we would distinguish is in terms of the accuracy of those representations. Some broadly externalist, indeed broadly reliabilist, approach has to be right when it comes to evaluating the visual systems of different animals.

Things are a little more complicated when it comes to evaluating individual visual beliefs of different animals, but it is still clear that we will use externalist considerations. So imagine we are looking for standards for evaluating particular visual beliefs of again fairly basic animals. One very crude externalist standard we might use is that a belief is good iff it is true. Alternatively, we might say that the belief is good iff the process that produces it satisfied some externalist standard, e.g. it is generally reliable. Or we might, in a way, combine these and say that the belief is good iff it amounts to knowledge, incorporating both the truth and reliability standards. It is not clear which of these is best. Nor is it even clear which, if any, animals without sophisticated cognitive systems can be properly said to have perceptual *beliefs*. (I won't pretend to be able to evaluate the conceptual and empirical considerations that have been brought to bear on this question.) But what is implausible is to say that these animals have beliefs, and the relevant epistemic standards for evaluating these beliefs are broadly internal.

---

[11] Jones attributes this view to Scotus and Ockham, and quotes Pedro Fonseca as saying almost explicitly this in his commentary on Aristotle's *Metaphysics*.

This matters to debates about the justificatory standards for *our* beliefs because we too have perceptual beliefs. And the way we form perceptual beliefs isn't that different from the way simple animals do. (If the representations of input processes are beliefs, then it doesn't differ in *any* significant way.) When we form beliefs in ways that resemble those simple believers, most notably when we form perceptual beliefs, we too are best evaluated using externalist standards. The quality of our visual beliefs, that is, seems to directly track the quality of our visual systems. And the quality of our visual system is sensitive to external matters. So the quality of our visual beliefs is sensitive to external matters.

On the other hand, when we reason, we are doing something quite different to what a simple animal can do. A belief that is the product of considered reflection should be assessed, inter alia, by assessing the standards of the reflection that produced it. To a first approximation, such a belief seems to be justified if it is well supported by reasons. Some reasoners will be in reasonable worlds, and their beliefs will be mostly true. Some reasoners will be in deceptive worlds, and many of their beliefs will be false. But this doesn't seem to change what we say about the quality of their reasoning. This, I take it, is the core intuition behind the New Evil Demon problem, that we'll address much more below.

So we're naturally led to a view where epistemic justification has a bifurcated structure. A belief that is the product of perception is justified iff the perception is reliable; a belief that is (or could have been) the product of reflection is justified iff it is well-supported by reasons. This position will remind many of Ernest Sosa's view that there is animal knowledge, and higher knowledge, or *scientia*. (See, for example Sosa 1991, 1997.) And the position is intentionally similar to Sosa's. But there is one crucial difference. On my view, there is just one kind of knowledge, and the two types of justification kick in depending on the kind of knower, or the kind of knowing, that's in question. If we simply form perceptual beliefs, without the possibility of reconsidering them (in a timely manner), then if all goes well, our beliefs are knowledge. Not some lesser grade of animal knowledge, but simply knowledge. To put it more bluntly, if you're an animal, knowledge just is animal knowledge. On the other hand, someone who has the capacity (and time) to reflect on their perceptions, and fails to do so even though they had good evidence that their perceptions were unreliable, does not have knowledge. Their indolence defeats their knowledge. Put more prosaically, the more you are capable of doing, the more that is expected of you.

8. The New Evil Demon Problem

The primary virtue of the above account, apart from its intuitive plausibility, is that it offers a satisfactory response to the New Evil Demon argument. The New Evil Demon argument concerns victims of deception who satisfy all the internal standards we can imagine for being a good epistemic agent. So they are always careful to avoid making fallacious inferences, they respect the canons of good inductive and statistical practice, they do not engage in wishful thinking, and so on. The core intuition of the New Evil Demon argument is that although these victims do not have knowledge (because their beliefs are false), they do have justified beliefs. Since the beliefs do not satisfy any plausible externalist criteria of justification, we conclude that no externalist criteria can be correct. The argument is set out by Stewart Cohen (1984).

A fairly common response is to note that even according to externalist epistemology there will be some favourable epistemic property that the victim's beliefs have, and this can explain our intuition that there is something epistemically praiseworthy about the victim's beliefs. My approach is a version of this, one that is invulnerable to recent criticisms of the move. (For both this response and the criticism to it, see Pryor 2001.) I'm going to call my approach the agency approach, because the core idea is that the victim of the demon is in some sense a good doxastic agent, in that all their exercises of doxastic agency are appropriate, although their perception is quite poor and this undermines their beliefs.

As was noted above, the quality of our visual beliefs is sensitive to external matters. This is true even for the clear-thinking victim of massive deception. Denying that the victim's visual beliefs are as good as ours is not at all implausible; indeed intuition strongly supports the idea that they are not as good. What they are as good at as we are is exercising their epistemic agency. That is to say, they are excellent epistemic agents. But since there is more to being a good believer than being a good epistemic agent, there is also for example the matter of being a good perceiver, they are not as good at believing as we are.

So the short version of my response to the New Evil Demon problem is this. There are two things we assess when evaluating someone's beliefs. We evaluate how good an epistemic agent they are. And we evaluate how good they are at getting evidence from the world. Even shorter, we evaluate both their collection and processing of evidence. Externalist standards for evidence collection are very plausible, as is made clear when we consider creatures that do little more than collect evidence. The intuitions that the New Evil Demon argument draws on come from considering how we process evidence. When we consider beliefs that are the products of agency, such as beliefs that can only be arrived at by extensive reflection, we naturally consider the quality of the agency that led to those beliefs. In that respect a victim might do as well as we do, or even better. But that is no threat to the externalist conclusion that they are not, all things considered, as good at believing as we are.

As I mentioned earlier, this is similar to a familiar response to the argument that James Pryor considers and rejects. He considers someone who says that what is in common to us and the clear-thinking victim is that we are both epistemically blameless. The objection he considers says that the intuitions behind the argument come from confusing this notion of being blameless with the more general notion of being justified. This is similar to my idea that the victim might be a good epistemic agent while still arriving at unjustified beliefs because they are so bad at evidence collection. But Pryor argues that this kind of deontological approach cannot capture all of the intuitions around the problem.

Pryor considers three victims of massive deception. Victim A uses all sorts of faulty reasoning practices to form beliefs, practices that A could, if they were more careful, could see were faulty. Victim B was badly 'brought up', so although they use methods that are subtly fallacious, there is no way we could expect B to notice these mistakes. Victim C is our paradigm of good reasoning, though of course C still has mostly false beliefs because all of their apparent perceptions are misleading. Pryor says that both B and C are epistemically blameless; C because they are a perfect reasoner and B because they cannot be blamed for their epistemic flaws. But we intuit that C is better, in some epistemic respects, than B. So there is some internalist friendly kind of evaluation that is stronger than being blameless. Pryor suggests that it might be *being justified*, which he takes to be an internalist but non-deontological concept.

The agency approach has several resources that might be brought to bear on this case. For one thing, even sticking to deontological concepts we can make some distinctions between B and C. We can, in particular, say that C is epistemically *praiseworthy* in ways that B is not. Even if B cannot be blamed for their flaws, C can be praised for not exemplifying those flaws. It is consistent with the agency approach to say that C can be praised for many of their epistemic practices while saying that, sadly, most of C's beliefs are unjustified because they are based on faulty evidence, or on merely apparent evidence.

The merits of this kind of approach can be brought out by considering how we judge agents who are misled about the nature of the good. Many philosophers think that it is far from obvious which character traits are virtues and which are vices. Any particular example is bound to be controversial, but I think it should be uncontroversial that there are some such examples. So I'll assume that, as Simon Keller (2005) suggests, it is true but unobvious that patriotism is not a virtue but a vice.

Now consider three agents D, E and F. D takes patriotism to extremes, developing a quite hostile strand of nationalism, which leads to unprovoked attacks on non-compatriots. E is brought up to be patriotic, and lives this way without acting with any particular hostility to foreigners. F is brought up the same way, but comes to realise that patriotism is not at all virtuous, and comes to live according to purely cosmopolitan norms. Now it is natural to say that D is blameworthy in a way that E and F are not. As long as it seems implausible to blame E for not working through the careful philosophical arguments that tell against following patriotic norms, we shouldn't blame E for being somewhat patriotic. But it is also natural to say that F is a better agent than either D or E. That is because F exemplifies a virtue, cosmopolitanism, that D and E do not, and does not exemplify a vice, patriotism, that D and E do exemplify. F is in this way praiseworthy, while D and E are not.

This rather strongly suggests that when agents are misled about norms, a gap will open up between blamelessness and praiseworthiness. We can say that Pryor's victim C is a better epistemic agent than A or B, because they are praiseworthy in a way that A and B are not. And we can say this even though we do not say that B is blameworthy and we do not say that being a good epistemic agent is all there is to being a good believer.

The proposal here obviously bears some similarity to the proposal by Kent Bach (1985), Mylan Engel (1992) and especially Clayton Littlejohn (manuscript) that we solve the New Evil Demon problem by saying it is possible that *S* is justified in believing that *p* while *S*'s belief that *p* is not justified. The thought then is that the victim might be justified in believing as they do, while their beliefs are not justified. And the same thing goes in general for agent's with a poor tie to reality. As I said, the agency approach moves along similar lines, but I do not rely on quite this distinction. The attraction of their position seems to come from identifying *S* with their epistemic agency, or perhaps with the part of their body that is the ground of their epistemic agency. It is hard to see what could motivate this. If *S* has several false beliefs because their eyesight is poor, they might not be a bad agent, but it is not because someone, or something, else went wrong. It is *their* eyesight after all that is poor. So it is a little odd to say that the person *S* is justified in their beliefs, even if we qualify this by saying the beliefs are not justified. Better, I think, to say that qua epistemic agent they didn't do anything wrong, and may indeed have done quite a lot right, but still that due to their poor eyesight they are not a particularly good believer.

9. Praise and Blame

As Pryor notes, many philosophers have thought that a deontological conception of justification supports an internalist theory of justification. I rather think that's mistaken, and that at least one common deontological understanding of what justification is entails a very strong kind of externalism. This is probably a reason to not adopt that deontological understanding.

Assume, for reductio, that S's belief that *p* is justified iff S is blameless in believing that *p*. I'll call this principle J=B to note the close connection it posits between justification and blamelessness. Alston (1988) seems to identify the deontological conception of justification with J=B, or at least to slide between the two when offering critiques. But one of Alston's own examples, the 'culturally isolated tribesman', suggests a principle that can be used to pull these two ideas apart. The example, along with Pryor's three brains case, suggests that A1 is true.

A1. It is possible for S to have a justified but false belief that her belief in *p* is justified.

A1 is a special instance of the principle that justification does not entail truth. Some externalists about justification will want to reject the general principle, but all internalists (and indeed most externalists) will accept it. Now some may think that the general principle is right, but that beliefs about what we are justified in believing are special, and if they are justified they are true. But such an exception seems intolerably ad hoc. If we can have false but justified beliefs about some things, then presumably we can have false but justified beliefs about our evidence, since in principle our evidence could be practically anything. So the following situation seems possible; indeed it seems likely that something of this form happens frequently in real life. S has a false but justified belief that *e* is part of her evidence. S knows both that anyone with evidence *e* is justified in believing *p* in the absence of defeaters, and that there are no defeaters present. So S comes to believe, quite reasonably, that she is justified in believing that *p*. But S does not have this evidence, and in fact all of her evidence points towards ~*p*.[12] So it is false that she is justified in believing *p*.

The following principle seems to be a reasonable principle concerning blameless inference.

A2. If S blamelessly believes that she is justified in believing that *p*, and on the basis of that belief comes to believe that *p*, then she is blameless in believing that *p*.

This is just a principle of transfer of blameworthiness. The quite natural thought is that you do not become blameworthy by inferring from *I'm justified in believing p* to *p*. This inference is clearly not necessarily truth-preserving, but that's not a constraint on inferences that transfer blameworthiness, since not all ampliative inferences are blameworthy. (Indeed, many are praiseworthy.) And it is hard to imagine a less blameworthy ampliative inference schema than this one.

I'm now going to derive a contradiction from A1, A2 and J=B, and a clearly consistent set of assumptions about a possible case of belief.

1. S justifiedly, but falsely, believes that she is justified in believing *p*. (Assumption - A1)
2. On the basis of this belief, S comes to believe that *p*. (Assumption)
3. S blamelessly believes that she is justified in believing that *p*. (1, J=B)
4. S blamelessly believes that *p*. (2, 3, A2)
5. S is justified in believing that *p*. (4, J=B)
6. It is false that S is justified in believing that *p*. (1)

One of A1, A2 and J=B has to go. If you accept J=B, I think it has got to be A1, since A2 is extremely plausible. But A1 only fails if we accept quite a strong externalist principle of justification, namely that justification entails truth. More precisely, we're led to the view that justification entails truth when it comes to propositions about our own justification. But as we saw above, that pretty directly implies that justification entails truth when it comes to propositions about our own evidence. And, on the plausible assumption that evidence can be practically anything, that leads to there being a very wide range of cases where justification entails truth. So J=B entails this strong form of externalism.

---

[12] I'm assuming here that evidence of evidence need not be evidence. This seems likely to be true. In Bayesian terms, something can raise the probability of *e*, while lowering the probability of *p*, even though the probability of *p* given *e* is greater than the probability of *p*. Bayesian models aren't fully general, but usually things that are possible in Bayesian models are possible in real life.

This does not mean that internalists cannot accept a deontological conception of justification. But the kind of deontological conception of justification that is left standing by this argument is quite different to J=B, and I think to existing deontological conceptions of justification. Here's what it would look like. First, we say that a belief's being justified is not a matter of it being *blameless*, but a matter of it being in a certain way *praiseworthy*. Second, we say that the inference from *I'm justified in believing that p* to *p* is not praiseworthy if the premise is false. So if we tried to run the above argument against J=P (the premise that justified beliefs are praiseworthy) it would fail at step 4. So anyone who wants to hold that justification is (even in large part) deontological, and wants to accept that justification can come apart from truth, should hold that justification is a kind of praiseworthiness, not a kind of blamelessness.

References

Adams, Robert Merrihew (1985) "Involuntary Sins" *Philosophical Review* 94: 3–31.

Alston, William (1988) "The Deontological Conception of Epistemic Justification" *Philosophical Perspectives* 2: 115-52.

Bach, Kent (1985) "A Rationale for Reliabilism" *The Monist* 68: 246–263.

Braddon-Mitchell, David and Frank Jackson (2007) *The Philosophy of Mind and Cognition*, second edition, Oxford: Blackwell.

Cohen, Stewart (1984) "Justification and Truth" *Philosophical Studies* 46: 279–295.

Cottingham, John (2002) "Descartes and the Voluntariness of Belief" *The Monist* 85: 343–360.

Davidson, Donald (1963) "Actions, Reasons and Causes" *The Journal of Philosophy* 60: 685-700

Descartes, René (1641/1996) *Meditations on First Philosophy*, tr. John Cottingham, Cambridge: Cambridge University Press.

Descartes, René (1644/1965) *The Principles of Philosophy*, tr. John Veitch, London: Dent

Engel, Mylan (1992) "Personal and Doxastic Justification in Epistemology", *Philosophical Studies* 67: 133-50.

Fodor, Jerry (1983) *The Modularity of Mind*, Cambridge, MA: MIT Press.

Gendler, Tamar Szabó (2000) "The Puzzle of Imaginative Resistance" *The Journal of Philosophy* 97: 55-81.

Gilbert, Daniel T. (1991) "How mental systems believe" *American Psychologist* 46, 107–119.

Gilbert, Daniel T., Douglas S. Krull, and Patrick S. Malone (1990) "Unbelieving the unbelievable: Some problems in the rejection of false information" *Journal of Personality and Social Psychology*, 59: 601-13.

Gilbert, Daniel T., Romin W. Tafarodi and Patrick S. Malone (1993) "You Can't Not Believe Everything You Read" *Journal of Personality and Social Psychology* 65: 221-233

Ginet, Carl (1985) “Contra Reliabilism” *Philosophical Studies* 68: 175-87.

Ginet, Carl (2001) “Deciding to Believe” in Steup (2001), pp 63-76.

Hasson, Uri, Joseph P. Simmons and Alexander Todorov (2006) “Believe It or Not: On the Possibility of Suspending Belief” *Psychological Science* 2006: 566-71.

Heller, Mark (2000) “Hobartian Voluntarism: Grounding a Deontological Conception of Epistemological Justification” *Pacific Philosophical Quarterly* 81: 130-141.

Hieronymi, Pamela (forthcoming) “Responsibility for Believing” *Synthese*, in press.

Holton, Richard (1999) “Intention and Weakness of Will” *Journal of Philosophy*, 96: 241-62.

Holton, Richard (2003) “How is Strength of Will Possible?” in Stroud and Tappolet (eds.) 2003, pp. 39-67.

Holton, Richard (2004) “Rational Resolve” *Philosophical Review* 113: 507-35.

Holton, Richard and Stephen Shute (forthcoming) “Self-Control in the Modern Provocation Defence” *Oxford Journal of Legal Studies*

Jones, Matthew L. (2006) *The Good Life in the Scientific Revolution: Descartes, Pascal, Leibniz and the Cultivation of Virtue*, Chicago: University of Chicago Press.

Keller, Simon (2005) “Patriotism as Bad Faith” *Ethics* 115: 563-592.

Kennett, Jeanette and Michael Smith (1993) “Philosophy and Commonsense: The Case of Weakness of Will” in *Philosophy in Mind: The Place of Philosophy in the Study of Mind* edited by Michaelis Michael and John O’Leary-Hawthorne, Dordrecht: Kluwer Academic Publishers, pp.141-157

Kennett, Jeanette and Michael Smith (1996) “Frog and Toad Lose Control” *Analysis*, 56: 63-73.

Littlejohn, Clayton (ms) “The Reliabilist’s Demon”

Owens, David (2000) *Reason Without Freedom: The Problem of Epistemic Responsibility*, New York: Routledge.

Pryor, James (2001) “Recent Highlights of Epistemology” *The British Journal for the Philosophy of Science* 52: 95–124.

Ryan, Sharon (2003) “Doxastic Compatibilism and the Ethics of Belief” *Philosophical Studies* 114: 47-79.

Ryle, Gilbert (1949) *The Concept of Mind*, New York: Barnes and Noble.

Shah, Nishi (2002) “Clearing Space for Doxastic Voluntarism” *The Monist* 85: 436-45.

Smith, Angela M (2005) “Responsibility for Attitudes: Activity and Passivity in Mental Life” *Ethics* 115: 236–71.

Smith, Michael (1997) “A Theory of Freedom and Responsibility” in *Ethics and Practical Reason* edited by Garrett Cullity and Berys Gaut, Oxford: Oxford University Press, pp. 293-319.

Smith, Michael (2003) “Rational Capacities” in Stroud and Tappolet (eds) 2003, pp 17-38.

Sosa, Ernest (1991) *Knowledge in Perspective*, Cambridge: Cambridge University Press.

Sosa, Ernest (1997) “Reflective Knowledge in the Best Circles” *Journal of Philosophy* 94: 410-30.

Steup, Matthias (2000) “Doxastic Voluntarism and Epistemic Deontology” *Acta Analytica* 15: 25-56.

Steup, Matthias (ed) (2001) *Knowledge, Truth and Duty*, New York: Oxford University Press.

Steup, Matthias (forthcoming) “Doxastic Freedom” *Synthese*, forthcoming.

Stich, Stephen (1978) “Beliefs and Sub-Doxastic States” *Philosophy of Science*, 45: 499-518.

Stroud, Sarah and Christine Tappolet (eds) (2003) *Weakness of Will and Varieties of Practical Irrationality*, Oxford: Oxford University Press, 2003

Watson, Gary (1977) “Skepticism about Weakness of Will” *Philosophical Review* 86:316-339.

Weatherson, Brian (2005) “Can We Do Without Pragmatic Encroachment?” *Philosophical Perspectives* 19: 417-43

Williams, Bernard (1973) “Deciding to Believe” in his *Problems of the Self*, Cambridge: Cambridge University Press.

Williamson, Timothy (2000) *Knowledge and Its Limits*, Oxford: Oxford University Press.