Philosophical Books Vol. 48 No. 2 April 2007 pp. 136–149

IN DEFENCE OF *ETHICS AND THE A PRIORI*: A REPLY TO ENOCH, HIERONYMI, AND TANNENBAUM

MICHAEL SMITH
Princeton University

Let me begin by thanking my commentators for their challenging responses to the essays reprinted in *Ethics and the A Priori* (hereafter EAP). They have given me a lot to think about, so much that I find myself with the practical problem of having to choose what to address and what not to address in the space available. Since it would be so easy to overlook or downplay the more serious difficulties, and to focus instead on the less serious; and since, if that were something I was inclined to do, I doubtless wouldn't represent myself to myself as doing it (!); let me begin with the blanket acknowledgement that the rather selective comments that follow are bound to be inadequate as they stand. If serious difficulties remain unaddressed, my only hope is that what I say will give at least some indication of where I think a fuller response might begin.

1. Reply to David Enoch

David Enoch's disagreements with me in his 'Rationality, Coherence, and Convergence' are about as deep and complete as disagreements can be. I will focus on his two central criticisms of the dispositional theory of value.

My proposal, you will recall, is that states of affairs in some possible world that is up for evaluation—let's call this the *evaluated* world—are desirable in virtue of the fact that my ideally rational self, who is just me in the nearest possible world in which I have a maximally informed and coherent desire set—let's call this the *evaluating* world—would desire that those states of affairs obtain in the evaluated world. The existence of value thus presupposes, I say, a partial convergence in the desires of our ideally rational selves: our counterparts in their evaluating worlds. Absent such a convergence, the desires we each have in our evaluating worlds would themselves seem too arbitrary to be of any normative significance.

As Enoch notes, I flip-flop in the essays about whether there are values, so understood. This is because, over the years, my conviction that our ideally rational selves would converge in their desires has waxed and waned.[1] Though this is not the place to explain why, it seems to me that it is a virtue of the proposed analysis of value that it permits such an equivocal attitude to

1. Most recently see my 'Is that all there is?' in *The Journal of Ethics*, 10 (2006), pp. 75–106.

Philosophical Books Vol. 48 No. 2

the existence of value, not a vice. Any analysis of value that didn't make it plain why the existence of value is the sort of thing about which creatures like us might have this kind of flip-flop attitude would seem to me unable to play a crucial role that an analysis of value has to play, which is to tell us something about—to use a rather old-fashioned term—the *human condition*. For the fact is that we do have a flip-flop attitude to the existence of value itself. This is why we agonize in the way we do, going back and forth about not just what the meaning of life is, but also about whether our lives are the sorts of thing that could have a meaning. I am therefore happy that my account supports this flip-flop attitude.

One of Enoch's two main complaints is related to this flip-flop attitude, for he thinks that one side of that flip-flop—the side where I find myself convinced that values *do* exist—is patently absurd. No one could take seriously the existence of value, understood in the way I propose, according to Enoch, because the required convergence is incredible.

> There are, it seems to me, infinitely many coherent sets of beliefs and desires. Perhaps Smith's own beliefs and desires comprise one such set. Perhaps the null set (or perhaps a set with many beliefs but no desires) is another, for where is the incoherence there? And there seems nothing in the very idea of coherence to exclude infinitely many other coherent sets of beliefs and desires. How is it, then, that all rational agents converge on the same set? Isn't this an amazing miracle? Surely, it cries out for explanation. And absent such an explanation it will be too much to believe. (p. 106 above)

Since it seems to me that an analysis of value must explain what it is that those of us who do believe in the existence of value believe in, this criticism strikes at the very heart of my proposal. However the criticism seems to me to fail because it assumes an account of what it is for a set of beliefs and desires to be coherent quite different from the one that I have in mind (for an account of what I have in mind see, for example, EAP pp. 312–314). Let me explain.

It should be generally agreed that there are all sorts of principles of rationality covering the relations our desires stand in to each other and to various beliefs. This is crucial to an understanding of the dispositional theory, for when I say that something is desirable if and only if and because we would desire that thing if we had a maximally informed and coherent desire set, what I mean thereby to be talking about is a desire set that best conforms to *all* such principles of rationality. Talk of coherence is, in other words, simply a *catch-all* term, one which allows us to talk of all of the principles of rationality that there are governing relations among desires and beliefs without being specific about what these principles are. Moreover, and happily, it allows us to do this even when we are unable to *say* what these principles of rationality are because we don't know what they are. We may, for example, be unsure whether some candidate principle is a principle of rationality or not. Even so, if that principle is a principle of rationality, then we can refer,

Philosophical Books Vol. 48 No. 2

inter alia, to the pattern of relations among our desires and beliefs that it mandates when we talk of a maximally coherent desire and belief set. Enoch may complain that this is a misuse of the word 'coherence.' But, misuse or not, this is how I use it.

But though we may not be able to say in advance what these principles of rationality are, we can say in advance something about the form that these principles of rationality might take, as the forms are the familiar ones much discussed by philosophers. They might, for example, take any of the following forms (in what follows 'RR' means 'Reason requires that':)

R1: RR (If someone has an intrinsic desire that *p* and a belief that he can bring about *p* by bringing about *q*, then he has an instrumental desire that he brings about *q*).
R2: RR (If someone has an intrinsic desire that *p*, and an intrinsic desire that *q*, and an intrinsic desire that *r*, and if the objects of the desires that *p* and *q* and *r* cannot be distinguished from each other and from the object of the desire that *s* without making an arbitrary distinction, then she has an intrinsic desire that *s*).
R3: RR (If someone has an intrinsic desire that *p*, then *p* is suitably universal).
R4: ∃p∃q RR (If someone believes that *p* then she has an intrinsic desire that *q*).

R1 is a generalization of the familiar means-end principle;[2] R2 is the sort of principle Derek Parfit appeals to when he criticizes Tuesday Indifference;[3] R3 is the sort of principle that Kantians think govern our maxims;[4] and R4 is the kind of principle that theorists like Tim Scanlon are committed to believing in insofar as they insist that there are reasons for desiring.[5] The question whether the desires of our ideally rational selves—understood, now, to be selves whose desires and beliefs are maximally informed and coherent—would converge thus turns on whether principles like these might underwrite such a convergence. Let me say a little about this.

Focus on R4. As I said, R4 is the kind of principle that Scanlon is committed to believing in insofar as he claims that there are reasons for desiring. This is because the fact that there are such reasons, if indeed there are, entails the possibility of reasoning oneself into having the requisite desires on the basis of one's grasp of those reasons: that is, on the basis of the belief that those reasons obtain. The situation, in other words, is much the same as when there exist certain reasons for believing. For in that case too the fact that there exist such reasons entails the possibility of reasoning oneself into having that belief on the basis of one's grasp of those reasons: that is, on the basis of believing

2. See my 'Instrumental Desires, Instrumental Rationality' in *Proceedings of the Aristotelian Society, Supplementary Volume* 78 (2004), pp. 93–109.
3. See Derek Parfit, *Reasons and Persons* (Clarendon Press, 1984).
4. See, for example, Christine Korsgaard, *The Sources of Normativity* (Cambridge University Press, 1996).
5. Thomas M. Scanlon, *What We Owe to Each Other* (Harvard University Press, 1998).

Philosophical Books Vol. 48 No. 2

that those reasons obtain. But the possibility of such a reasoning process is simply the assumption that there is a principle of rationality like R4.

What might such a reasoning process look like and how might it underwrite a convergence in the desires of our ideally rational selves? Consider a simple example. Suppose that the state of affairs in which people are as happy as possible is intrinsically desirable. In that case, according to the analysis I propose, it follows that we would all converge on the intrinsic desire that people are as happy as possible if we had a maximally informed and coherent desire set. But now it should be clear how, given what I have in mind by the coherence, this could turn out to be so. For there may be a reason provided by something intrinsic to that state of affairs itself—presumably the fact that happiness has the nature that it has—which provides us all with a reason to have an intrinsic desire that the state of affairs in which people are as happy as possible obtains. But if there is such a reason then that entails the existence of a principle of rationality governing the relationship between this belief and our desires, a principle with the form of R4. More specifically, it entails that there is a principle of rationality like R4 where *p* is 'Happiness has the nature that it has' and *q* is 'People are as happy as possible'. And if there is such a principle of rationality then, since our maximally informed and coherent selves will know about the reason—they are, after all, *informed*, so they will believe that happiness has the nature that it has—and since they will reason from this belief to the desire that people be as happy as possible in the way R4 demands—they are, after all, *coherent*, so their desires and beliefs will conform to all of the principles of rationality that there are, including R4—so it follows that our maximally informed and coherent selves will all converge on an intrinsic desire that people be as happy as possible.

Note that this discussion of the way in which principles of rationality like R4 might underwrite a convergence in desires really is just illustrative, for I might equally have focused on the way in which principles of rationality like R3 could do the same job. According to Kantians, for example, the only desires that are suitably universal are those whose realization would leave intact the ability of the desirer himself to exercise his rational capacities, and would similarly leave intact the ability of others to exercise their rational capacities. The idea is thus that we are able to derive a principle of the form of R4 from a principle of the form of R3, and then the reasoning proceeds as before. In this case, however, the R3 style principle is fundamental. Here too, then, we see how a principle of rationality might so constrain the desires that we would have if we had a maximally informed and coherent desire set that a convergence in our desires would result.

As Enoch points out, I say in various essays that in order to figure out whether values exist we have little alternative but to engage in various normative arguments and see where they lead. There is therefore, I say, a sense in which drawing substantive conclusions in meta-ethics—actually forming beliefs about the *existence* of values, as distinct from beliefs about the appropriate *analysis* of value—is beholden to normative ethics. It is, I hope, now clear why I say this. For if the existence of value would be clinched by the existence of a principle of rationality that has the form of R4 like the one just discussed,

Philosophical Books Vol. 48 No. 2

or a principle with the form of R4 derived from a principle of rationality that has the form of R3, then one normative dispute that looks clearly relevant to settling questions about the existence of value is a normative dispute about whether or not there are such principles of rationality. Answering this question requires us to give the arguments for and against the existence of such principles of rationality and see where those arguments lead. My analytic suggestions were never meant to be a substitute for this kind of substantive normative argument.

Enoch's second main objection to the dispositional theory is that there is slippage in the idea I employ of an *ideally rational advisor*. My proposal, you will recall, is that states of affairs in some possible world up for evaluation—the *evaluated* world—are desirable in virtue of the fact that my ideally rational self, whom we are to imagine off in the nearest possible world in which I have a maximally informed and coherent desire set—the *evaluating* world—would desire that those states of affairs obtain in the evaluated world. I sometimes put the point by saying that we can think of my ideally rational self as an advisor, telling me how the evaluated world is to be. What's desirable is that the evaluated world be the way my ideally rational self, over there in the evaluating world, advises me it is to be. But of course talk of advice is a mere heuristic. What's crucial is not that my ideally rational self gives me advice—since communication across possible worlds is impossible, he does no such thing—but rather that my ideally rational self has certain desires about the ways things are to be in the evaluated world.

Enoch objects that I make the dispositional theory look more plausible than it really is by connecting it to the idea of what an ideally rational advisor would advise.

> [I]t seems to me rather commonsensical that my ideally rational advisor—in the sense needed for the connection between desirability and his advice—must satisfy much more than coherence (and full non-normative information). It seems to me that he must also have the right views on many matters, including many normative matters. In particular, he must know what really is desirable, or at least what really is desirable for me. And indeed, his desire-set need not be the minimally revised (that is, informed and coherent) set of desires, the informed and coherent set most resembling my original one. If I have some desires for things that are not desirable (for me), my ideal advisor should simply not have those desires, and shouldn't have them precisely because they are not desirable (rather than them being undesirable because he doesn't desire them). And if the minimally revised set of my desires still fails to include desires for all and only what is desirable, my ideally rational advisor's desire-set should be different from my minimally revised one. These are, it seems to me, the job requirements for an ideally rational advisor, if desirability is to be understood in terms of his advice. But, of course, on such an understanding it doesn't seem that a dispositional theory of value is still in play, for the ideal advisor's dispositions are now a function of what is really, independently, desirable, and not the other way around. (p. 100f. above)

Philosophical Books Vol. 48 No. 2

But the previous discussion of the way in which various principles of rationality might underwrite a convergence in our desires should make us doubt this.

Enoch wants us to imagine that, on the one hand, there are facts about which states of affairs are desirable, facts to which my ideally rational advisor is sensitive in giving advice. But he also wants us to imagine, on the other hand, that the features of those states of affairs in virtue of which they are desirable provide my ideally rational advisor with no reason at all to desire that those states of affairs obtain. To imagine otherwise, after all, is to imagine, contrary to Enoch, that there is indeed a relevant principle of rationality governing the formation of my ideally rational advisor's desires, a principle with the form of R4, either a fundamental such principle or one derived from an R3 style principle. It is therefore crucial to Enoch's objection that the features of states of affairs that are desirable provide no such reason. But the idea that my ideally rational advisor advises that some state of affairs is to obtain, and yet that the features of those states of affairs in virtue of which they are desirable provide him with no reason at all to desire those states of affairs obtain is hard to take seriously. Why would he advise that those states of affairs obtain, as opposed to any other, as opposed to being totally indifferent, if the features of those states of affairs provided him with no reason to desire that they obtain?

But now suppose, contrary to Enoch, that the features of the states of affairs in virtue of which my ideally rational advisor advises that those states of affairs are to obtain provide him with a reason to desire that they obtain. In that case what should we say about the desirability of such states of affairs? The obvious thing to say, contrary to Enoch, is that the desirability of those states of affairs isn't an independent feature to which my ideally rational advisor is sensitive. Desirability is better understood as being constituted by the fact that my ideally rational advisor would desire that those states of affairs obtain if he formed his desires about them in conformity with all of the principles of rationality that there are. In this way we see how the mere fact that the desires of our ideally rational advisors are maximally informed and coherent—where, remember, coherence must be understood in the way I intend, where this may include conformity to principles of rationality like R3 or R4—may indeed suffice to forge the required connection between the desirability of states of affairs and the advice of our ideally rational advisors. Indeed, it is hard to imagine on what basis an ideally rational advisor could give advice unless some such story were true.

2. Reply to Pamela Hieronymi

In 'Rational Capacities as a Condition on Blame' Pamela Hieronymi takes issue with my attempt to explain the kernel of truth in the claim that, to be responsible, an agent must have the ability to do otherwise. The approach I favour sees this as a consequence of the fact that, when an agent is responsible for something that happens, this is because the occurrence of

Philosophical Books Vol. 48 No. 2

that thing is suitably explained either by the fact that he exercised (when he is due praise) or by the fact that he failed to exercise (when he is due blame) some relevant rational capacity: roughly speaking, what we hold agents responsible for are those things that happen as a consequence of their responding or failing to respond to reasons to the extent that they have the capacity to do so.

In order to flesh out this story I attempt to describe the precise similarities and differences between various very simple cases in which someone fails to respond to reasons: cases in which an agent, John, fails to come up with the correct answer to some philosophical question which he is asked. My idea, in focusing on these simple cases, is that the best way to work out what to say about more complicated moral cases is by starting from these simple cases and seeing what we need to add to them to build up a moral case. (Hieronymi is therefore right that I hold "failing to exercise a capacity is a necessary but not a sufficient condition for legitimate blame" (p. 117 above).) In particular, I try to describe the differences between Blanking John—this is an agent who has, but fails to exercise, the capacity to come up with the correct answer to some philosophical question he was asked on some occasion—and Ignorant John, an agent who does not possess the capacity to come up with the correct answer to such a question.

Though the details of my proposal are rather complicated, the crucial feature, as Hieronymi notes, is that when an agent has certain rational capacities, that fact about him turns out to be constituted by a certain characteristic pattern of similarities and differences in the way the agent is in actuality and the way he is in those possible worlds in which he is and is not responsive to the reasons he has. Here, accordingly, we find the kernel of truth in the claim that responsibility requires the ability to do otherwise: responsibility for some outcome in actuality requires a certain characteristic pattern of similarity and difference between the way the agent is in actuality and the way he is in a range of nearby possible worlds in which he is differently responsive to reasons.

Hieronymi admits that the general approach I favour—to repeat, this is the approach that sees the 'could have done otherwise' condition on responsibility as a consequence of the fact that responsibility presupposes that that for which we are responsible is a consequence of our exercise or failure to exercise certain rational capacities that we possess—is "one many people find natural" (p. 111 above). She is sceptical about this approach, however, as she thinks that it requires us to conceive of moral capacities as *exceptional*.

> It is worth noting, at the start, that this line of thought separates our moral capacities from capacities of other sorts. Typically, our capacities develop as demands are put upon us to exercise them well—beyond our current ability. We usually learn to think more clearly, write more smoothly, run more swiftly, plan more effectively, or sing more beautifully because someone (often oneself) expects better of us. Moreover, in many areas of adult life—in one's career, in one's role as teacher or parent, in one's position as chair or as second tenor—the demands one is under remain insensitive to one's own particular shortcomings; one's capacities develop as one tries to

Philosophical Books Vol. 48 No. 2

meet them. If, in contrast, we cannot be put under genuine moral demand until we are already able to satisfy it, the development and exercise of our moral capacities would be, in this respect, fairly exceptional. (p. 111f. above)

But it seems to me that she is wrong about this.

Focus on the details of a particular case. Imagine my singing teacher demands more of me than I am able to deliver, and, as a result, that I acquire the ability to sing better. What happens in such a case would seem to be this. By asking me to sing even better than I can at *t*, my singing teacher brings it about that I sing in some way or other at *t*—let's call this way W—where singing in way W at *t* causes me to acquire the ability to sing even better at a subsequent time *t′* than I could at *t*. But now it seems that my singing teacher really makes two quite different kinds of demands of me. She explicitly demands that I sing better than I can at *t* and she implicitly demands that I sing in way W at *t*. True enough, at *t* I cannot sing better than I can at *t*, but, *ex hypothesi*, at *t* I can sing in way W at *t*. Moreover it seems crucial that this is so, for the implicit demand that I sing in way W is something for which my singing teacher can and presumably does hold me responsible in the sense that Hieronymi admits most people find natural: that is, I put myself in a position where blame is warranted if I fail to sing in way W and where praise is warranted if I succeed, precisely because I can do what my singing teacher has implicitly demanded that I do.

What this suggests, to me at any rate, is that Hieronymi equivocates on 'demand'. Let's call the thing that my singing teacher asks me to do implicitly—to sing in way W, something that I can do—a *demand simpliciter*, and let's call the thing she asks me to do explicitly—to sing better than I can—an *aspirational demand*. What Hieronymi observes is that there are many cases in which we make aspirational demands of people, where aspirational demands make no presuppositions about our capacities. Let's agree with that. But the fact seems to be that we can only be subject to aspirational demands by being subject, at least implicitly, to demands simpliciter as well. Moreover, to repeat, it also seems that, at *t*, we can only be subject to a demand simpliciter to do something at *t* that we can do at *t*. Hieronymi's examples of aspirational demands would thus seem to confirm rather than to disconfirm the natural view that we can only be subject to demands—that it is to say, demands simpliciter—that we can satisfy.

The suggestion that this view "separates our moral capacities from capacities of other sorts" (p. 111 above) would therefore seem to be mistaken. The natural approach is about demands simpliciter, not aspirational demands, and so it quite rightly makes no claim at all about how we might develop our moral capacities. Those of us who go along with the natural approach can therefore accept much that Hieronymi says about aspirational demands. We can accept what she says because it is not in conflict with, and indeed would seem to presuppose, the more natural approach to demands simpliciter.

Focus now on demands simpliciter. As I have said, I attempt to illuminate such demands by describing the difference between Blanking John, who has

Philosophical Books Vol. 48 No. 2

but fails to exercise the capacity to give the correct answer to a philosophical question he is asked, and Ignorant John, who doesn't possess that capacity. Though Hieronymi concedes that the "crucial points" I make in providing this description "seem important and correct" (p. 115 above), she does not think that I have thereby succeeded in describing a situation in which it was *up to Blanking John* to give that answer.

> Smith has shown that Blanking John had the capacity to have answered correctly, in the sense that, in a host of nearby possible worlds, he would have answered correctly. So Blanking John *could* have answered correctly, one might object, only in the sense that Blanking John *might* have answered correctly. Smith has secured a metaphysical possibility, here—in fact, a strong one: in very many nearby possible worlds, Blanking John answers correctly—but he has not thereby secured any sense in which it is *up to John* whether he thinks of the right answer, when asked. Possibility—even strong possibility—does not amount to control.
>
> Suppose I have a heart attack. It may well be that, in a host of possible worlds, in circumstances similar to the one I am in, I do not have a heart attack. Further, the fact that I do not have a heart attack in those worlds may be explained by the underlying structure of my cardio-vascular system. Thus, I have the capacity, in Smith's sense, to have not suffered the heart attack in my current circumstances. The truth of this claim does nothing to show that it was up to me whether I had a heart attack; neither does it show that I had any opportunity to avoid the heart attack. (pp. 118–119 above)

But the objection doesn't work.

To have a capacity, in the sense I describe, is to display a certain pattern of sensitivity in the formation of beliefs and desires to the rational relations that those beliefs and desires stand in to other beliefs and desires and to available evidence. It is thus crucial that the capacity ranges over *intentional states*. Nor should this be surprising, given that what I am attempting to describe is not just any old capacity, but a *rational* capacity. Having the capacity that Hieronymi imagines, by contrast—the capacity not to have a heart attack—plainly isn't a matter of displaying such a sensitivity. True enough, this capacity is what it is in virtue of "the underlying structure of my cardio-vascular system". But it would be absurd to suggest that this structure is what it is in virtue of the fact that it underwrites exactly the rational relations between intentional states characteristic of a rational capacity. The upshot—completely unsurprisingly, or so I hope—is that Hieronymi is wrong to suggest that "I have the capacity, in Smith's sense, to have not suffered the heart attack in my current circumstances".

As I have said, in my view it suffices for an agent to be in control of some outcome that that outcome is suitably explained either by that agent's possession and exercise of some rational capacity, or by her possession and failure to exercise some such capacity. Illumination of the concept of control therefore requires that we come up with an analysis of our rational capacities and their exercise. What we need to be on the lookout for, in providing such an

Philosophical Books Vol. 48 No. 2

analysis, is whether at some point we find ourselves having to posit something like libertarian free will. Hieronymi's suspicion is that we will (p. 121 above). Her main argument for that conclusion is that my own analysis, which makes no such posit, cannot distinguish between the exercise of a rational capacity and the capacity not to have a heart attack, something we've just seen to be false. But she has another argument as well.

> [I]f blanking John is culpable for his blanking, the explanation of his failure must run out at the fact that he simply failed to exercise some capacity. His failure to exercise the relevant capacity is itself inexplicable. But for being inexplicable it seems also uncontrollable, and therefore unavoidable. (p. 119 above)

I have to confess that I do feel the force of this objection, but I think we should resist it. To say that John's failure to think of the right answer is inexplicable is simply to acknowledge that reason explanations have run out at this point. *Ex hypothesi*, John didn't have a reason for failing to think of the right answer, he simply failed to think of the right answer despite the fact that he had the capacity to do so. But to my mind *that very fact*—the fact that he had the capacity to do so—entails that the exercise was under his control in the relevant sense. So I think we should deny that its being inexplicable entails that it is uncontrollable.

If we can provide an analysis of rational capacities and their exercise of the kind I suggest then we can use that analysis to provide an account of moral responsibility with demands simpliciter, and hence control, at center stage. Hieronymi disagrees with this whole approach, as she prefers a more normative interpretation of what it is for an outcome to be up to an agent, one in which the kinds of aspirational demand discussed earlier play a foundational role. But, for the reasons given, it seems to me that such aspirational demands cannot play a foundational role: the aspirational demands Hieronymi herself describes presuppose the existence of demands simpliciter, and hence need to be built up out of an account of rational capacities and their exercise of the kind that I attempt to provide. Hieronymi is a nimble opponent, however, so she will doubtless disagree with this in turn on the basis of some more or less compelling reason, so perhaps I should just admit that there is doubtless much more that needs to be said on both sides about this issue.

3. Reply to Julie Tannenbaum

In 'The "Should" of Full Practical Reason' Julie Tannenbaum seems happy enough to go along with what I say about the relationship between desirability and reasons for action, or, in her terminology, 'should' claims. Her worry is rather that I do not make enough distinctions among such claims.

According to Tannenbaum, we need to introduce a new kind of 'should' that applies to actions, a 'should' different both from the 'should' of instrumental

Philosophical Books Vol. 48 No. 2

rationality and the 'should' of what she calls "full practical reason" (hereafter 'should$_{FPR}$'). The latter, in my terms, is the 'should' that expresses what my fully rational self—which, to repeat, is just me in the evaluating world where I have a fully informed and coherent desire set—would want myself to do in the evaluated world, where—and this will be important in what follows—the wants of my fully rational self in the evaluating world are restricted to those he has for outcomes in the evaluated world that can be realized by actions that are among the options I have in the evaluated world.[6]

Tannenbaum brings out the need for this extra 'should' by means of an example.

> As I stand in the store deciding which cigarettes to buy, my friend says, "You should buy the low tar cigarettes." What is the nature of the 'should' in my friend's claim? Consider how we assess the truth of my friend's claim. It is shielded from certain sorts of facts and grounded on, or sensitive to, others. If a bystander were to object by saying, "No you shouldn't, because you shouldn't be smoking at all" the bystander would have misunderstood the nature of my friend's claim. The truth assessment of my friend's claim is shielded from certain sorts of facts—for instance from the fact that I should desire to be healthy more than I desire the pleasure of smoking—and is grounded on other facts: that I desire to smoke and that I should desire to be healthy, where the latter desire is treated not as a competitor with my desire to smoke, which I believe it properly is, but rather as a constraint on how I pursue the satisfaction of my desire to smoke. (p. 127f. above)

She insists that this 'should' expresses neither the 'should' of instrumental rationality nor should$_{FPR}$. So what does it express?

The first point I want to insist upon is that we can certainly imagine a variation on this case in which the 'should' is 'should$_{FPR}$'. In deciding whether or not this is so the question we need to ask is whether refraining from smoking is an *option* for me. Here, accordingly, we see the importance of the restriction emphasized above: the wants of my fully rational self that are relevant to my reasons for action are those he has for outcomes of *options* that I have in the evaluated world. If not smoking isn't an option for me—if, say, I'm addicted in some very strong sense to cigarettes—then it simply isn't true that I should$_{FPR}$ not smoke. Rather I should bring about the best of the outcomes I *can* bring about: perhaps I should$_{FPR}$ smoke low tar cigarettes.

Tannenbaum anticipates this response in her discussion of another case, one in which I say a squash player who cannot control his anger and would smash his opponent in the face if he got near him after being defeated, should$_{FPR}$ leave the court without shaking his opponent's hand, notwithstanding the fact,

6. My ideally rational self may, of course, desire that I do all sorts of things that aren't among my options. My doing these things would be good, but I would have no reason to do these things. For more on this distinction see Philip Pettit and Michael Smith, 'External Reasons', in Cynthia Macdonald and Graham Macdonald (eds.), *McDowell and His Critics* (Blackwell, 2006) pp. 140–168.

Philosophical Books Vol. 48 No. 2

assuming it to be a fact, that his fully rational self would prefer him to control his anger and shake his opponent's hand (p. 126 above). For here too it seems to me important that the squash player's controlling his anger and shaking his opponent's hand isn't an option for him. Tannenbaum rejects this response (pp. 126ff. above). She thinks, correctly, that the restriction is motivated by the thought that 'ought'—and hence 'should$_{FPR}$'—implies 'can'. But, she says, cases of negligence show that there is no such straightforward move from 'ought' to 'can'.

Consider the drunk driver who runs over a pedestrian. He should$_{FPR}$ have stopped notwithstanding the fact that he couldn't have stopped *at the time*: that is, despite the fact that it wasn't an option. He should have stopped because he shouldn't have been drunk, and if he hadn't been drunk, then he would have stopped. But it seems to me that we can say all of this without saying, what in any event sounds wrong to me, that he should$_{FPR}$ have stopped *at the time* and hence without denying 'ought' implies 'can'. The crucial point is that drinking earlier *was* one of his options. This means that, consistently with 'ought' implies 'can', we can say he should$_{FPR}$ have stopped in one sense (that is, he should$_{FPR}$ have done something that he could have done earlier that would have resulted in his stopping later) even though it isn't the case that he should$_{FPR}$ have stopped in another (that is, even though it isn't the case that he should$_{FPR}$ have done something that he could have done later that would have resulted in his stopping later). He is thus responsible because he could and should$_{FPR}$ not have drunk so much *earlier*, which would have had the consequence that he stopped later. But since he couldn't stop at the later time, it isn't the case that he should$_{FPR}$ have stopped *at that time*.

But now let's suppose that Tannebaum is right and that it isn't true that I should$_{FPR}$ smoke low tar cigarettes. Suppose, in other words, that I do have the option of not smoking, and hence that this is what I should$_{FPR}$ do. In that case how are we to understand my friend's claim that I should smoke low tar cigarettes? Do we need to introduce a whole new sort of 'should' claim, beyond those I have been talking about? Not really. For remember something I said at the very beginning. In my view facts about reasons for action are themselves analyzable into two components: what I have reason to do is a matter of, first, what my options are, and second, the desirability of the outcomes of my taking those options, where desirability is understood in terms of the dispositional theory. These resources are sufficient to provide us with an understanding of the 'should' claims that interest Tannenbaum, or so it seems to me.

Note that, if reasons do factor into the two components suggested, then we can readily distinguish between what I have *most reason* to do, what I have *next most reason* to do, what I have *next most reason after that* to do, and so on. I have most reason to do that thing, among my options, that produces the *best outcome*; I have next most reason to do that thing, among my options, that produces the *next best outcome*; I have next most reason after that to do that thing, among my options, that produces the *next best outcome after that*; and so on. And note as well that the dispositional theory tells us how to understand this ranking of outcomes. The best outcome is the one I would *most want* if I

Philosophical Books Vol. 48 No. 2

had a maximally informed and coherent desire set; the next best outcome is the one I would *next most want*; the next best after that is the one I would *next most want after that*; and so on.

Applying all of this to the smoking case, the idea is that even though I should$_{FPR}$ not smoke because that is the option, among those options I have, that has the outcome that my fully rational self most wants, I should in another sense—one which takes that option out of the equation—smoke low tar cigarettes because this is the option, among the options that now I have left, that has the next best outcome: in other words, it is the option whose outcome is that which my fully rational self next most wants. And so we could go on defining all the 'should's we could ever want. I am therefore very happy to agree with Tannenbaum that we need many more 'should's. But the materials with which to construct them seem to me to be the very ones that the theory of reasons for action that we get out of the dispositional theory of value makes available.

In the final section of her paper Tannenbaum takes issue with my account of what we should$_{FPR}$ do. In particular, she points out that we do not hold agents responsible for doing what they should$_{FPR}$ do, understood in the way I suggest, because what they should$_{FPR}$ do is fixed by facts to which they may have no epistemic access: after all, I say that what we should$_{FPR}$ do is a function of what I would want if I had a maximally *informed* and coherent desire set. She gives the following example.

> Imagine a hiker on a path where there is no history of anyone ever having been injured in this area. Nevertheless, there is, for the first time, an unconscious injured person off trail and out of sight or hearing. The hiker has the physical capacity to save this injured person (she can physically reach the injured person and properly tie a tourniquet). But the fact that someone is injured off trail is not rationally accessible to the hiker, and so she cannot rationally come to the conclusion that he should save the person. If she were to go off trail looking for potential people to save, she would discover the injured person. But any course of reasoning that leads a mere day hiker (as opposed to a ranger) to the conclusion 'I'll wander off trail looking for injured persons in case there is one' will not be rational. She would be unreasonably endangering herself. Since there is no appropriate line of reasoning that would lead her to believe that there is someone injured off trail, then there is no appropriate (non-defective) line of reasoning that would lead her to believe that she should help the person off trail. And so, through no fault of his own, saving the person is not something she rationally can do. For this reason, I believe that saving this person is not something the hiker should$_{FPR}$ do. (p. 133 above)

It might be thought that this is a particularly damaging criticism given my earlier emphasis on the fact that 'ought', and hence 'should$_{FPR}$', implies 'can'. But in fact I think the criticism isn't damaging at all when you think it through.

What I have insisted upon is that we should$_{FPR}$ do only those things that are options for us. The first question to ask is therefore whether, in the hiker

Philosophical Books Vol. 48 No. 2

case, finding the injured person is one of the hiker's options. The answer, let's agree, is that it is. The next question to ask, however, is whether failing to take an option that we should$_{FPR}$ take suffices for responsibility. And the answer to this question must surely be that it isn't sufficient. What agents are responsible for, as I've tried to make clear in my discussion of Hieronymi's paper, are those outcomes which are appropriately explained by those agents exercising or failing to exercise their rational capacities. As applied to actions, this means that agents can only be responsible for those actions which they perform or omit as a result of their exercising or failing to exercise their rational capacities. But by this criterion it follows that the hiker isn't responsible for failing to find the injured person. He isn't responsible because, for the very reasons Tannenbaum gives, no exercise of his rational capacities could have led him to desire to do *that*. Being an option is thus a necessary, but not a sufficient condition, for responsibility. What Tannenbaum's example shows thus isn't what she thinks it shows. Saving the injured person is indeed something that the hiker should$_{FPR}$ do, it just isn't something which he is responsible for failing to do.

Philosophical Books Vol. 48 No. 2