Converging on values

Donald C. Hubin

In *The Moral Problem*, Michael Smith not only analyses what he takes to be the central problem of moral philosophy, he sketches a solution to that problem. The problem is, briefly, that of reconciling our belief in the objectivity of moral statements with our belief in the motivational force of such statements given the acceptance of a theory that grounds motivation in the subjective conative states of agents. The sketch of a solution depends crucially on the distinction between a motivating reason and a normative reason. While the former are based on the desires of the agent, the latter are determined by desires that a fully rational agent would have concerning the behaviour of the actual agent. Smith seeks to show how reasons based on such considerations can have just the sort of connection to actual motivation that normative reasons do and how such reasons can be seen as objective, in an appropriate sense.

In addition to being objective, Smith believes, normative reasons, if there are any, must be non-relative. Thus, while accepting a Humean account of motivating reasons, Smith rejects a Humean account of normative reasons. To be more precise, Smith believes that, while an agent's motivating reasons are dependent on the agent's desires or some other subjective, contingent, conative state of the agent, his normative reasons are not. Indeed, they are not relativized to the agent at all despite the fact that their status as normative reasons for that agent depends on the fact that his ideally rational self would desire that he act on them. Relativity is avoided, Smith believes, because all rational agents would, through a process of rational deliberation and correction of false beliefs, converge on at least some desires concerning how to act in a given set of circumstances. Call this 'the convergence hypothesis'. Not only does Smith believe that the convergence hypothesis is true, he believes that were it not true, there would be no normative reasons for action (1994: 173).

The idea of non-relative normative reasons that have the proper rational authority over us has been an attractive one to philosophers. And many have sought to ground such reasons on the desires idealized agents would have. I have elsewhere (1996) attacked the claim that reasons based on such desires have the proper normative authority. Here, I want to consider only the claim of non-relativity. While Smith doesn't claim to have shown that there *are* non-relative reasons with appropriate rational authority, he does claim that we should be optimistic and this implies optimism about the truth of the convergence hypothesis. One does not, of course, refute

Analysis 59.4, October 1999, pp. 355–61.

optimism. But one can throw a little cold water on it. And, while the role of pessimist is hardly as uplifting as that of optimist, I shall play it as convincingly as possible.

It would be an easy trick – too easy to be worth doing – to defend the convergence hypothesis if one claimed that the concept of a rational agent or that of rational deliberation directly imposes strong, substantive requirements on desires. For example, if one held that it was simply, on the face of it, irrational to fail to desire that suffering be diminished, one could easily conclude that all rational agents would, by a process or rational deliberation, converge on having a desire that suffering be diminished. Then, given Smith's conception of a normative reason, one could conclude that there is a normative reason to lessen suffering.

Smith, though, employs no simple sleight of hand. His conception of the idealizing conditions under which convergence is to be hoped for are seemingly quite noncontroversial. Following Bernard Williams, Smith requires that:

(i) the agent must have no false beliefs
(ii) the agent must have all relevant true beliefs
(iii) the agent must deliberate correctly. (1994: 156ff.)[1]

Clearly, condition (iii) does most of the hard work of forcing convergence, if convergence there is to be. Correcting false beliefs and introducing true beliefs can obviously produce convergence on derivative desires. So, two agents might have different desires about welfare policy and, as a result of the removal of false beliefs and the introduction of relevant true ones concerning the actual effects of various policies, that disagreement might disappear. How far this sort of process can go in producing convergence of desire among real human beings is an issue for debate. But it seems unlikely to resolve conative conflict where the disagreement is fundamental – where the desires in question are intrinsic rather than instrumental.

Some would put this much more strongly, saying that doxastic correction alone can never generate a change in nonderivative desires. This is one way in which, according to its critics, the Humean view limits the power of reason to alter an agent's desires. Smith may well agree with this, but it seems to me to be wrong. Elimination of error and acquisition of true beliefs can lead one to see the causes, effects and nature of a nonderivative

[1] Of course, nothing in this area is really non-controversial. For one thing, the significance of the convergence hypothesis depends on the assumption there is a real distinction to be made between belief and desire, which some have denied, and that desirability cannot be smuggled into the belief states mentioned in (i) and (ii). Smith believes that judgements of value – of the desirability of a thing – are cognitive states of mind. But, of course, when we idealize agents for his project, we do not assume that they have true beliefs about value. This would be blatantly circular.

desire. As a result of other conative aspects of the agent's psychology, this knowledge can generate a reason, and a motive, to alter the nonderivative desire. For example, if I found that my desire to engage in dangerous sports had the effect of frightening my loved ones, I might be led to alter that desire because of my stronger desire not to produce this effect. This is not because the thrill seeking was in some way derivative from some other desire. A second example – of the tail-swallowing sort philosophers enjoy – is the following. Suppose that I find that my having a nonderivative desire for my own happiness thwarts that happiness because it leads others to think of me as selfish and, hence, not to trust me. I may well analyse the situation correctly and alter my nonderivative desire for my own happiness precisely so as to satisfy it.

Even recognizing how the correction of doxastic states can alter nonderivative as well as derivative desires, though, the first two conditions offer little reason for optimism about the convergence hypothesis. This becomes more obvious when we realize that the hypothesis is not an empirical one about the convergence of desires among *human beings* when their beliefs are corrected and they deliberate properly. Such convergence might depend crucially, not only on the process of deliberation, but on contingent facts about the fundamental desires common to human beings. Smith's convergence hypothesis must assert that there would be convergence in the desires of all conceivable rational beings who have no false beliefs and all relevant true ones and who deliberate correctly. As Smith puts it:

> Which desires *I* would end up with, after engaging in such a process, thus in no way depends on what *my* actual desires are to begin with. Reason itself determines the content of our fully rational desires, not the arbitrary fact that we have the actual desires that we have. (1994: 173)

While doxastic correction might provide a more powerful converging force among humans than is sometimes suspected, recognition of these effects should leave us far from optimistic about the truth of the convergence hypothesis. We can certainly imagine agents whose fundamental desires are radically different from our own in ways that would be untouched by removing false beliefs and instilling relevant true ones. And this leaves us relying on the third condition, which requires rational deliberation, to ensure the sort of convergence that will result in the nonrelative judgements of value that can ground normative reasons.

With respect to this condition, also, we need to avoid trivialization. Correct deliberation cannot be defined in terms of its outcome. Indeed, 'correct deliberation' must be analogous to what Rawls (1971: 88f) calls 'pure procedural justice' in the sense that the correctness of the outcome

must be defined entirely in terms of the correctness of the procedure employed to arrive at it. If there were a definition of 'correct desire' that was independent of the correct deliberative procedure, then Smith would have no need to discuss correct deliberation in order to fill out his theory. Value could be defined directly in terms of correct desires and we could simply understand normative reasons as those based on value in this sense. The point of Smith's approach is that value is 'constructed' from the notions of correct deliberation by rational, error-free agents.

So, hopes for the truth of the convergence hypothesis hang heavily on correct deliberation. Unfortunately, Smith has little to say about the concept of correct deliberation. Clearly and noncontroversially, it must include means/ends reasoning. Beyond that, Smith criticizes Williams' account as making too much of imagination's role and ignoring the far more important role of 'systematic justification': 'by far the most important way in which we create new and destroy old underived desires when we deliberate is by trying to find out whether our desires are *systematically justifiable*' (1994: 158–59).

The process of seeking a systematic justification of our desires involves 'trying to integrate the object of that desire into a more coherent and unified desiderative profile and evaluative outlook' (1994: 159); we are seeking the sort of 'reflective equilibrium' described by John Rawls (1971) and developed further by Normal Daniels (1979). While Smith attempts to show why this sort of coherentist mode of justification is applicable to our desires, as well as our beliefs, his discussion doesn't add much to the familiar concept of reflective equilibrium, and many worries remain. For example, one might wonder if a set of desires that is unified and systematized in the way Smith's process requires is in any sense more justified than one that is not. Suppose I desire to eat strawberries and desire to eat raspberries and desire to eat oranges, but desire not to eat blackberries and desire not to eat tangerines. Would I have a more justified set of desires if I altered them so that I desired to eat the three types of berries and desired not to eat either citrus fruit? What error am I making when I have the original desires that should be corrected by a process of correct deliberation? Coherence seems rather more clearly a virtue when we talk about values and desirability than when we talk of desires. But, for Smith, desirability and value is just desire that exists in agents whose cognitive states are corrected and who deliberate correctly. So, he seems required to say what he means by calling one set of desires more justified than another and to show why a unified and systematized set of desires is more justified.

Here, though, I will not pursue this criticism further because our focus is on the convergence hypothesis. What I want to ask is whether there is anything in the description of correct deliberation, as Smith understands it,

that offers serious hope for the truth of that hypothesis. I don't think there is; worse, I think we find there fertile ground for pessimism.

In the first place, there is no reason to suppose that anything in the process of achieving reflective equilibrium will produce convergence in the cases of radically disparate initial desiderative states. The sort of case that seems congenial to the convergence hypothesis might be one like this. Two people start out with desires for the flourishing of their friends and without any similar desire for that of strangers. In the process of trying to unify and systematize their desires, they find no difference between their friends and strangers – at least none they are willing to assert as relevant. As a result of their deliberations, they both come to desire the flourishing of all.

Whatever the success of correct deliberation in producing convergence of desires in these sorts of cases, there seems to be no reason to suppose that it will produce convergence between one who desires for its own sake the flourishing of others and one who desires for its own sake the suffering and death of others. Correct the doxastic states fully. Let the people deliberate at length about the most systematic, unified desiderative structure that might emerge from these starting desires. I see no reason for believing that they would converge in their desires; rather, it seems they would remain forever divided.

Maybe there are no individuals who desire for its own sake the suffering and death of others. Or, maybe if there are such people, they also have other desires that would provide a fulcrum to move them toward the sort of convergence that more socially adjusted people would attain. But this is convergence as a result of a lucky fact about human nature. It is not the process of correct deliberation that guarantees convergence but the nonrational aspects of the human beings engaging in it. And, in order to address the sort of relativism Smith worries about, the convergence hypothesis must be more robust than this.

Secondly, to make us share his optimism about the convergence hypothesis, Smith must give us some reason to be pessimistic about the *divergence* hypothesis. This hypothesis, which is much weaker than the convergence hypothesis, holds that for at least some conceivable agents the idealizing process Smith describes would increase the disparity between their desire sets. People beginning with only slightly divergent sets of desires might, it would seem, be pulled in dramatically different directions by the process of correcting cognitive errors and deliberating correctly. I may begin by desiring a slightly more heavily graduated income tax than you do. It may well happen, as we become aware of the relevant facts, lose our mistaken beliefs and deliberate as Smith supposes, that we are led far from these initial desires in opposite directions. I may become deeply desirous of a radically egalitarian society and you of a meritocratic or *laissez faire*

society. As a result, our desires about taxation policy may differ even more radically than prior to the process of deliberation.

Finally, different, but equally unifying and systematizing, general desires may exist for a given desiderative profile. It seems antecedently to be true that the same initial set of desires could be unified and systematized in two equally good, though quite different ways. Just as many theories can account for any finite set of observations, so the particular desires that we begin with are compatible with many different more general desires that might unify and systematize them. And, remembering that we cannot appeal to a standard of correctness for desires that is external to the process of ideal deliberation, there is no reason to believe that there will always be one that does so best.

Some terminology might clarify the point: Let's call one set of desires, D_n, an 'admissible successor' to another, D_m, just in case an agent, beginning with the desire set D_m and committing no rational error, can arrive at D_n by a process consisting only of the removal of false beliefs, the introduction of relevant true beliefs and what Smith describes as correct deliberation. What I'll call 'the underdetermination thesis' holds that there is at least one set of desires such that it has at least two admissible successors. I think the underdetermination thesis is extremely plausible – indeed, almost certain to be true – given how Smith characterizes the process of correct deliberation. If so, then it is not the case that there is, antecedent to the actual process of deliberation, a fact of the matter about what one's ideally rational self would desire – at least not a fact of the matter that is independent of arbitrary psychological features of the agents. And when those arbitrary features are taken into account, real agents with identical desire sets might well have rational idealizations with arbitrarily divergent desire sets despite the fact that all of them have deliberated correctly from corrected doxastic states.

The underdetermination thesis seems extremely plausible; there seems to be no reason to be sceptical of the truth of the divergence hypothesis; and, nothing Smith describes in the process of ideal deliberation even hints at a basis for optimism about convergence of radically divergent desire sets. These factors combine to make me extremely pessimistic about the prospects for convergence.

It is worth stressing, again, that the pessimism I am pushing does not concern the convergence of desires of actual human beings under the process Smith describes. (Though I'm very pessimistic about that, as well.) Even if we were to observe evidence of such convergence, it would provide little hope for the sort Smith needs to lay relativity to rest. Convergence of desires among actual humans engaging in rational deliberation may well be the result of the *humanity* of the agents rather than their *rationality.* Our

humanity may be crucial to the explanation of convergence in one of at least two ways. First, as was already mentioned, we may begin with desiderative profiles that are similar as a result of non-rational aspects of our human nature. Second, it may be that we are, in virtue of our humanity, inclined toward arriving at normative consensus (Gibbard 1990). So, while there may be good reasons for us to adjust our desires so that they converge with those of others with whom we interact, this may not be the result of there being good reasons for arriving at one convergence point rather than another. I take it neither of these explanations will serve Smith's anti-relativistic program.

The implications of the pessimism I've been peddling for Smith's theory are significant. He believes that the very concept of a normative reason is dependent on non-relativity. If the rest of his story is correct, then, the falsity of the convergence hypothesis would entail not merely the absence of non-relative normative reasons, but the complete absence of normative reasons. A high price to pay, I think. But, of course, a price that can be avoided by recognizing that there can be normative reasons that are relative to an agent's conative set. Given the grounds for pessimism about the convergence hypothesis, I think the recognition of relative normative reasons looks more and more attractive.[2]

The Ohio State University
Columbus, OH 43210, USA
hubin.1@osu.edu

References

Hubin, Donald. 1996. Hypothetical motivation. *Noûs* 30: 31–54.
Daniels, Norman. 1979. Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy* 76: 256–82.
Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgement.* Cambridge, MA: Harvard University Press.
Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Harvard University Press.
Smith, Michael. 1994. *The Moral Problem.* Cambridge: Blackwell.

[2] For very helpful comments on an earlier draft of this paper, I am grateful to Justin D'Arms and Andy Wallace.