Do the desires of rational agents converge?

David Sobel

Michael Smith, in his impressive book *The Moral Problem*,[1] claims that (1) 'convergence in the hypothetical desires of fully rational creatures is required for the truth of normative reason claims' (173) and (2) we have reason to 'have some confidence [… that] there will be a convergence in our desires under conditions of full rationality.' (187). These claims play a crucial role in supporting Smith's distinctive and interesting meta-ethical position. However, I will here ignore the context of Smith's broader ambitions and focus on these two theses. I will argue that Smith does not give us adequate justification for either claim.

This paper will have four sections. In §1 I will explain Smith's conception of fully rational deliberation. In §2 I discuss what it would be for desires to converge as Smith suggests. In §3 I consider Smith's conceptual claim that such convergence is necessary for there to be normative reasons at all. Finally, I address Smith's reasons for optimism about convergence. The first three sections consider Smith's first thesis above while the fourth section considers his second thesis.

1. The plausibility of the claim that the desires of all agents will converge after proper deliberation hinges crucially on how one characterizes such deliberation. One could simply claim that a person only counts as having deliberated properly if she reaches certain approved conclusions. This path would assure Smith's first thesis at the cost of invoking a substantive, non-proceduralist conception of proper deliberation. The interest in Smith's claims stems from his willingness to invoke an understanding of proper deliberation which is not conceptually tied to the deliberator arriving at any particular motivations.

In this paper I will say that a conception of normative reasons for action is Humean iff

(1) it claims that one's rational desires give one normative reasons (hereafter, reasons), and
(2) it invokes a proceduralist notion of proper deliberation in which what makes a desire rational is that it would be arrived at if one successfully followed the procedure, and
(3) this procedure is specified such that, at least in principle, any particular option could be rationally desired after successfully

[1] Smith, 1994. All otherwise unattributed references are to this work.

Analysis 59.3, July 1999, pp. 137–47.

following the procedure, and

(4) it is relativist in the sense that different agents might arrive in the right way at different desires and hence have divergent reasons.

This last condition shows that, for the Humean, the normativity for A of the desires that A would arrive at in the approved way is not undermined should other agents arrive in the approved way at other desires. Smith clearly rejects the fourth thesis. He must also be rejecting the first or second thesis. Perhaps he would want to say that successfully following the procedure makes one's resultant desires rational, but that (contra thesis 1) rational desires only give one reasons when all rational agents converge in their desires. Alternatively, Smith might want to say that one's desires are only rational if everyone would reach them after proper deliberation (contra thesis 2).

Smith finds himself in substantial agreement with Bernard Williams's account of fully rational deliberation, complaining only that Williams's account requires supplementation.[2] (158) Thus Smith and Williams largely agree about what it is for an agent to arrive at a desire in the right way. Smith glosses Williams's position in this way: in order for a deliberator to counts as fully rational '(i) the agent must have no false beliefs, (ii) the agent must have all relevant true beliefs, (iii) the agent must deliberate correctly.' (156)

Conditions (i) and (ii) claim that we can only be fully rational deliberators if we get the facts right. Smith's requirement of true belief shows that he is not trying to capture the notion of proper deliberation given one's limited time and access to the truth. Thus while we might have been rational in light of the information available to us (because, for example, the scientific community had not discovered certain facts) or that it was reasonable for us to collect, we were not fully rational deliberators in Smith's sense for we did not know all the facts.

One sense of 'rational deliberation' is to make proper use of the information one has (or could reasonably be expected to get). Thus a person might have been rational in this sense to leave the building after the fire alarm went off even though it turns out that the building was not really on fire. Smith is clearly after another notion of rational deliberation, one perhaps well designed to determine what it is advisable to do or what will get us what we really want. Smith's targeted notion might be called 'ideally rational deliberation.' Several philosophers, myself included, have recently argued that this notion of ideally rational deliberation involves serious conceptual problems, but I want to focus on other issues here.[3]

Condition (iii) adds possibilities for practical reasoning such as these:

[2] Williams 1981 spells out his neo-Humean account of instrumental rationality.

the agent adequately exercises her imagination in considering the options available, the agent makes no mistakes of means-end reasoning, the agent engages in constitutive reasoning 'such as deciding what would make for an entertaining evening, granted that one wants entertainment.' (Williams 1981: 104) Smith's main additions to Williams's above criteria are that the agent is not suffering from depression, apathy, or other debilitating attitudes, and the agent has a concern that her overall desire pattern be coherent and justified. (158–61)

Crucially these possibilities for practical reason do not amount to the claim that it is criterial of proper deliberation that one reach a particular specified motivational profile.

2. Before considering Smith's arguments for his two theses we must better understand what we mean when we speak of the desires of all ideally rational agents converging. This is more complicated than it might seem.

Different things might be thought to constitute convergence. A weak understanding of convergence would be that all fully rational agents agree about every claim of this form: A has normative reason to O in circumstances C. If a Humean account of reasons were correct and this was accepted, then, presumably, we would all agree, when fully rational, about all such reason claims. In fact, the most likely way to achieve such agreement would be for all rational agents to agree on an account of practical rationality; that is, agree about what makes it the case that A has reason to O in circumstances C. This kind of agreement is compatible with different agents having arbitrarily divergent ideally rational desires. Call this 'convergence about each agent's reasons.' This, it turns out, is not the notion of convergence that Smith is pointing to. Or rather, such agreement would seem necessary but not sufficient for the kind of convergence that Smith has in mind.

Smith's understanding of convergence would be for all rational agents to converge on desires that have the same *de se* content. (169–70) If we both want you rather than me to get the larger slice of cake there is a sense in which we converge and a sense in which we do not. We both want the same state of the world to obtain, but we do not both want the cake for ourselves. If we both wanted cake for ourselves our desires would have the same *de se* content. A simple case of this sort of convergence would be if all fully rational agents desired only their own pleasure. In this case rational agents agree about what kind of things are desirable.

[3] See Velleman 1988, Sobel 1994, Rosati 1995, and Loeb 1995. Interestingly all these critiques are offered against the full information account of well-being rather than the remarkably comparable full information account of reasons for action that Williams and Smith are developing. These two theories, for all their similarity, seem to have developed and be discussed in isolation from each other.

Convergence in the objects of desire can be thicker or thinner. We might both want to listen to the Rolling Stones album *Exile on Main Street* very loud, or we might both want to listen to music. The former agreement in desire is thicker in that the shared object of appropriation is specified more specifically. It is likely that some very thin convergence can always be found between two people or perhaps between all rational agents. Perhaps all rational agents desire to be rational, to live a good life, or to avoid mistakes. If the convergence that Smith has in mind is convergence in the objects of desire, then it is crucial to know what level of generality he intends. Thick convergence among all rational agents would be very surprising and, perhaps, entail moral agreement, whereas very thin convergence would not be very interesting and would not entail moral agreement. Further, if we are not given a specific level of generality at which rational agents are supposed to converge, then one could worry that Smith could, in the face of any seeming disagreement, simply move in each such case to thinner and thinner areas of convergence. There could also be varying degrees of the extent of convergence. Rational agents might agree (at a fixed level of generality) about all, most, or just a few objects of desire. Smith's claim is that there is full convergence in desire amongst rational agents.

A problem for the thesis that rational agents will converge on the objects of desire thickly described is that on matters of mere tastes agreement does not seem forthcoming. It is not plausible that every rational deliberator will like chocolate ice cream more than vanilla or Seinfeld more than Friends. Our desires diverge over such issues. Yet in many cases our divergent desires seem for all the world informed in the relevant sense and well insulated from consistency or justificatory pressures from other areas of our motivational make-up. There are strong grounds for thinking such divergence in tastes will survive rational deliberation.

Smith concedes this point, but tries to show that this would not prevent the kind of full convergence he has in mind.[4] Smith admits that his preference for beer over wine provides him a reason to get beer rather than wine where our opposite preference provides us reason to get wine instead. Smith claims that this is no threat to full convergence because the preference is 'a relevant feature of our circumstances....' (171) This can be seen, Smith tells us, 'from the fact that I can quite happily agree with you that if I were in your circumstances – if I preferred wine to beer – then the fact

[4] Seemingly, Smith might have conceded that the case of mere tastes shows that there would not be full convergence on the objects of desire after fully rational deliberation, while continuing to insist that we have reason to be 'optimistic about the possibility of an agreement about what is right and wrong being reached under more idealized conditions of reflection and discussion.' (189)

that the local wine bar sells very good wine would constitute a reason for me to go there as well ….' (171) Smith continues: 'even if an agent's preferences may enter into a specification of the circumstances that she faces it may still be the case that whether or not she is rationally justified in taking her own preferences into account, and the way in which she is justified in taking them into account if she is, depends on whether fully rational agents would all converge on a desire which makes the preferences she has relevant to her choice.' (171)

Smith is here claiming that rational agents might, despite their divergent tastes, nonetheless converge on desires of the following form: if I were in circumstances C (which includes having a certain preference profile P), then I want myself to choose O. At first sight such convergence might appear to be nothing more than convergence on instrumental rationality itself. It appears that the thought is simply this: whatever desires I find myself having, I want myself to satisfy those desires. But things are not so simple.

Allan Gibbard notes that 'Ideals differ from tastes: I dislike spinach but think it a matter of taste; that means in part that although I dislike spinach I am willing to eat it if I like it. The norms I accept endorse eating it if I like it and not otherwise. I oppose cruelty unconditionally: I want myself not to be cruel even if, hypothetically, I should want to be.' (Gibbard, 1990: 167)

Thus, for those that hold ideals in Gibbard's sense (and moral claims seem paradigmatically to be such ideals), what one would want for oneself if one had an altered preference ranking need not simply be maximally to satisfy those altered desires. Smith's claim seems to be that all rational agent's might converge in their *de se* desires about what they would want for themselves if they were in a particular circumstance (which includes having a certain preference profile).

There are at least two different ways in which such convergence might be achieved. The first way would be for all rational agents to treat every preference as a matter of taste and never as an ideal. Such rational agents might agree that for any situation they find themselves in, they would want to maximally satisfy the preferences that they have in that situation. Convergence of this sort, I take it, would be compatible with a thoroughgoing Humean account of an agent's reasons for action. Thus I suppose it must not be this method of reaching convergence that Smith has in mind for he calls his proposal an 'anti-humean theory of normative reasons.' Smith, it would seem, must therefore have in mind a picture in which fully rational agents converge in their desires in a way that crucially involves them having ideals. How might such a picture go?

The second way such convergence might occur, in this case involving agents with ideals, would be for all rational agents to agree fully about what counts as a matter of mere taste and what counts as a matter of

ideals. It would not be important that the rational agents agree in their tastes, rather only that they agree about which of their preferences are to be treated as mere tastes. For where we treat our preferences as mere matters of taste, we presumably agree about what we want in situations in which we have this taste or that; namely, indulge our tastes as we find them. In this way Smith has found a nice way of acknowledging that fully rational agents would not converge on preferring chocolate ice cream to vanilla, yet still holding out the prospect that they might fully converge in their desires in an interesting and important sense.

It would seem however that the rational agents would have to converge on their ideals in order for the full kind of convergence in desires that Smith envisions to occur. When two agents disagree in their ideals, they will sometimes disagree in their wants for themselves in circumstances C (which includes having a certain preference profile P). Thus for the kind of convergence that Smith has in mind to occur all rational agents must both (1) agree about what counts as a mere matter of taste and what counts as an ideal and (2) agree in their ideals.

Imperfectly rational agents clearly disagree about what are matters of mere taste and what are ideals. Some want themselves to dress fashionably even in the situation in which they do not want to look fashionable, others treat this as a matter of taste. It must be claimed that this kind of disagreement would be extinguished upon becoming fully rational.

It is worth noting that for two people to agree in their ideals they must agree about significantly more than just morality. For as the above example of fashion shows, a person can hold a preference as a personal ideal even when they think others would not be immoral or irrational for failing to share that ideal. Gibbard called such ideals 'existential: it is a choice of what kind of person to be, in a fundamental way, come what might, which the chooser does not take to be dictated by considerations of rationality.' (*ibid*.: 168) Thus it would seem that to vindicate Smith's conception fully rational agent's would have to converge fully not only about morality, but also about personal ideals. This would seem to necessitate that the process of becoming fully rational extinguish all existential ideals, or at least make false the claim that the holder of the existential ideal makes: namely that commitment to this ideal is not dictated by rationality. And as holding false beliefs is supposed to be incompatible with being fully rational, it would seem that Smith must hold that existential ideals must go if convergence is to take place.

Let me here parenthetically acknowledge a clear and important virtue of Smith's attempt at an 'anti-Humean Theory of Normative Reasons.' It deals as successfully with matters of mere taste as does the Humean theory and this to my mind had been one of the clearest advantages of the

Humean account. Smith shows us how the anti-Humean can, perhaps effectively, avoid conceding this front.

3. Why must proper deliberation lead to convergence if we are to have reasons at all? Smith's conclusion here is remarkably strong. He claims the very concept of reasons demands convergence. Thus, apparently, the dominant Humean understanding of reason is, in a straightforward way, conceptually confused. I find Smith's clearest arguments for this to be the following:

> (1) such an assumption is necessary to underwrite the platitude that 'we are … potentially in agreement or disagreement with each other about what constitutes a reason and what doesn't.' (172), and
> (2) absent this sort of convergence I must see my reasons as being entirely dependent on the fact that my actual desires are as they are. But the shape of my actual desires (which, by hypothesis, will determine the shape of my desires after rational deliberation), Smith tells us, is an 'entirely arbitrary matter, one without any normative significance of its own.' (172)

I will consider these arguments in turn.

Consider first Smith's case that convergence is necessary to vindicate the thought that we can dispute reasons claims. He writes

> there is … a sense in which we can talk about rational justification or desirability *simpliciter*. When you and I talk about the reasons that there are for acting [in this sense], we are therefore talking about the same thing …. On the relative conception, however, matters are quite different. For in order to give a truth condition to the schematic claim 'It is desirable that *p* in *c*' we need to know from whose perspective the truth of the claim is to be assessed …. [Thus] if I say to you 'There is a reason for O-ing', and you deny this, we are therefore potentially talking about different things …. (166–67)

All of this is true. But it does not support either the claim that on the Humean view there could not be a common subject matter when we discuss reasons, or that 'rational justification [concerning reasons] is itself a relative matter; that really there is only rational-justification-relative-to-this-person or rational-justification-relative-to-that.' (171) In fact these two claims are false.

Smith claims that on Williams's Humean view when we speak of reasons 'it typically means $\text{reason}_{\text{me}}$ out of my mouth, $\text{reason}_{\text{you}}$ out of your's, $\text{reason}_{\text{her}}$ out of her's and so on.' (172) But, of course, we might both be talking about the reasons of the same person. We might be debating if A has reason to O in circumstances C. On the Humean account when we

debate whether A has reason to O in circumstances C we are debating what A's motivations would be like if she were deliberating properly. The Humean thus holds that we can be right or wrong in such claims and we have a common subject matter when we discuss them. Someone who claimed that a person had an intrinsic reason to O, when that person had nothing in her motivational set that O answered to, would just be wrong. Thus the Humean framework allows us to vindicate the platitude that we have a common subject matter when we speak of a particular agent's reasons. Perhaps it is the case that, on the Humean view, we cannot find a common subject matter for other kinds of reason claims. Likely the Humean cannot vindicate certain kinds of reason claims that we might have hoped to vindicate (e.g., that there are reasons, such as moral reasons, which apply to all rational agents as such). But even if this is granted there is no reason here to suppose, as Smith does, that the Humean is forced to abandon the very notion of reasons altogether.

Further, the way to justify reason claims such as A has reason to O in circumstances C is not relative to each individual. The Humean holds that there is a fact of the matter about this which is determined by the agent's motivations in counterfactual situations (which will admittedly be very difficult to determine) just as there is a fact of the matter about the effects of red wine on the body. And we justify believing the former sort of facts in the same non-person relative way that we justify the latter sort of facts.

Smith's second argument claims that unless desires converged amongst fully rational deliberators, one's rational desires would be determined by what one (perhaps idiosyncratically) happened to find attractive. Smith claims that if this were so one's rational desires would depend on 'an entirely arbitrary matter' and 'arbitrariness is precisely a feature of a consideration that tends to undermine any normative significance it might initially appear to have.' (172–73) Smith could mean one of two things. First, he could mean that as a psychological matter an agent cannot take her own desires seriously, cannot herself find them to be a source of reasons for her, if they are arbitrary in this way. This is clearly false. The prevalence of sincere Humeans demonstrates this. The Humean may be wrong about reasons for action, but it must be allowed that she really believes her theory.

The second interpretation of Smith's claim is that there is no justification for treating one's desires that are arbitrary in this way as creating reasons. But this is exactly what is at issue between the Humean and Smith. My rationally arbitrary preference (that is, a preference I could be rational without) for chocolate ice cream gives me a reason to choose it. What is incoherent about all my reasons being such? Smith has baldly claimed that there is a conceptual confusion in the notion of rationally arbitrary desires by themselves creating reasons. The predominant neo-Humean view holds

that this is precisely what creates reasons. Thus Smith will have to do better. I myself have trouble seeing anything conceptually confused or problematically arbitrary in the thought that my rationally optional projects and commitments create reasons for me and not for you.

Before moving on I want to briefly consider the plight of the agent that deliberates by successfully following the procedure that Smith recommends but unhappily discovers that other rational agents arrived at divergent desires. Smith tells us that this agent has no reason to satisfy her desires, in fact she has no reasons (recall I use 'reason' to refer to Smith's 'normative reason') at all for there are no reasons without convergence. Thus a lack of convergence, according to Smith, would make it the case that I had no reason to avoid pain or choose a flavour of ice cream I like rather than one I find disgusting. I find this suggestion difficult to take seriously. Does Smith really think the existence of such reasons is contingent on universal convergence?

Perhaps Smith should only claim that an agent has a reason to O iff all rational agents would converge on a desire for O in the relevant circumstances rather than claim that for an agent to have any reasons all rational agents must converge on all desires. I would still disagree with this claim, but it would not flaunt what I regard as the absurdities mentioned in the previous paragraph.

In some cases, like the preference for chocolate rather than vanilla ice cream, our bare liking provide us with a reason to choose one thing rather than another. Even if all rational agents do not converge in their desires, I like chocolate ice cream better than vanilla, and this is all I need to have a reason to choose it. It would seem that Smith must either claim that, in such cases, my getting chocolate ice cream does not benefit me more than getting vanilla or that I have no reason whatsoever in such a case to seek my benefit rather than my detriment. Neither claim strikes me as plausible.

4. Smith argues that convergence in desires amongst rational deliberators is quite plausible. Smith's main argument here is that 'the empirical fact that moral argument tends to elicit the agreement of our fellows gives us reason to believe that there will be a convergence in our desires under conditions of full rationality. For the best explanation of that tendency is our convergence upon a set of extremely unobvious *a priori* moral truths.' (187) Smith bolster this argument with three supplementary arguments: (1) often by focusing only on moral disagreement we ignore substantial agreement and the fact that 'we' share thick evaluative language which entails a fair degree of agreement in attitude, (2) although current moral disputes appear at times deadlocked, 'we must remember that in the past similarly entrenched disagreements were removed *inter alia* via a process

of moral argument' and (3) some moral disagreement certainly can be explained as clearly arising from lack of rational deliberation. (188)

But much of the historical moral argumentation that has actually managed to produce consensus has been (1) factually and logically imperfect, (2) addressed to those poorly positioned to object, (3) addressed to those who share substantial common moral vocabulary, moral education, and cultural identification, and (4) offered by those who are persuasive for reasons other than the cogency of their position. Any historical tendency towards convergence that results from some combination of these four causes (and no doubt others) will not constitute evidence that rational agents will converge in their moral views. Since such causes undoubtedly played a very significant role in producing convergence, we cannot simply point towards a general historical tendency towards convergence (even if there were one) and claim it constitutes inductive evidence for Smith's case. Thus, as we might have suspected, Smith's historical case about a tendency towards convergence will have to be genuinely historical. It will have to persuade us of the crucial role of facts, logic, and reason is explaining the history of convergence and the secondary role of force, guile, and a shared thick moral vocabulary.

The sort of convergence Smith holds out real hope for is convergence amongst people that start off as different in their desires, moral vocabulary, and moral education as we can imagine. The convergence in the objects of desire that Smith needs is not simply convergence amongst all actual humans (as if we could make ourselves right by killing dissenters) but rather amongst all possible agents no matter how initially divergent. But what is the history of convergence amongst agents with radically divergent desires, moral vocabularies, and moral educations when they are confronted with factually informed and logically impeccable argumentation? Certainly Smith will have to remind us of such cases. Smith needs a string of cases in which (1) convergence occurred, (2) it occurred for the right reasons, and (3) it occurred between radically different cultures.

Smith relies quite plainly and inappropriately on convergence within an extended community. Smith points to convergence within a community that shares a common moral vocabulary of thick evaluative language. (188) This commitment to a common thick evaluative language, Smith rightly points out, reveals considerable moral agreement within the community. (188) But just because this shared thick evaluative language presupposes and reveals considerable moral agreement, convergence amongst those that share thick evaluative language is of very limited use in demonstrating the power of rational argumentation to itself manufacture this agreement. Rational argumentation has some power to lead people who actually agree about morality to see that they agree. This is not

doubted. But Smith needs to show that history provides a good inductive case that all fully rational agents, no matter how initially divergent they are prior to becoming fully rational, will agree on moral matters.

Smith finds himself 'in substantial agreement with Derek Parfit' (214) when Parfit claims that

> Belief in God, or in many gods, prevented the free development of moral reasoning. Disbelief in God, openly admitted by a majority, is a very recent event, not yet completed. Because this event is so recent, Non-Religious Ethics is at a very early stage. We cannot yet predict whether, as in Mathematics, we will all reach agreement. (Parfit 1984: 454)

But betraying such hopes undermines Smith's inductive historical argument for it suggests that the factually informed deliberators will not predicate their moral opinions on religious authority. And surely an overwhelming amount of historical ethical argument that has produced convergence has relied on religious beliefs and authority. It might be true, as Parfit says, that absent such a religious basis we lack any reason to suppose that there will not be convergence. But surely we must therefore also admit that past moral practices do not provide sufficient inductive grounds to suppose that such agreement will be forthcoming.

Thus I conclude that Smith's arguments in support of the plausibility of extensive convergence are unconvincing and that he has failed to show that such convergence is necessary for the existence of reasons.[5]

Bowling Green State University
Bowling Green, OH 43403-0222
sobel@bgnet.bgsu.edu

References

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Harvard University Press.
Loeb, D. 1995. Full-information theories of individual good. *Social Theory and Practice* 21: 1–30.
Parfit, D. 1984. *Reasons and Persons*, Oxford University Press.
Rosati, C. 1995. Persons, perspectives, and full information accounts of the good. *Ethics* 105: 296–25.
Smith, M. 1994. *The Moral Problem*. Oxford University Press.
Sobel, D. 1994. Full information accounts of well-being. *Ethics* 104: 784–10.
Velleman, D. 1988. Brandt's Definition of 'Good'. *Philosophical Review* 97: 353–71.
Williams, B. 1981. Internal and external reasons. In his *Moral Luck*, Cambridge University Press: 101–13.

[5] I would like to thank David Copp, Michael Smith, Sara Worley and an anonymous referee for *Analysis* for their help with this paper.