

Michael Smith
Princeton University
msmith@princeton.edu

Original Scientific Paper

UDC 17.022.1

111.821

Received: September 26, 2024

Revised: October 06, 2024



Accepted: October 10, 2024

RETHINKING *THE MORAL PROBLEM*

Abstract

Are intrinsic desires subject to reasoned criticism, and if they are, what is about them that makes them subject to such criticism? It is argued that though the answer given to this question in *The Moral Problem* is wrong, a more promising answer can be found if we attend to the metaphysics of agency.

Keywords: Motivating reasons vs normative reasons · categorical reasons vs hypothetical reasons · intrinsic desire · criticism · convergence · agent · agency · ideal vs non-ideal · reflective equilibrium

It is thirty years since the publication of *The Moral Problem* (hereafter *TMP*). Though my views have changed in various ways since writing the book, to my mind they leave much of the argument intact. The central organizing problem in metaethics still seems to me to be that presented in the first chapter: that is, the *prima facie* inconsistency of the claim that morality is objective in a sense that rules out both non-cognitivism and metaethical relativism, the claim that there is some sort of necessary connection between moral judgements and motivation, and the claim that Hume was right about the nature of motivation. Moreover, I remain convinced that the solution to this problem lies in a version of moral rationalism of the kind introduced in the second chapter, a kind that is in turn committed to a dispositional theory of value similar to that argued for in the fifth chapter. Having said that, in what follows I will focus on some of the ways in which my views have changed since writing *TMP*, emphasizing those prompted by what I now think of as mistakes.

1. Folk metaphysics vs fundamental metaphysics

Perhaps the biggest change in my thinking since writing *TMP* concerns the metaphysical ambition of metaethics. Between the late 1970s and early 1990s my teachers included Frank Jackson and Simon Blackburn, and my colleagues

included David Lewis, Frank Jackson, and Philip Pettit. Following their lead, the most general question I took myself to be answering was 'What do we think the world is like and is it that way?', where this question could be raised about both fundamental and non-fundamental worldly features. The non-fundamental worldly features with which we're most familiar are folk features. Folk moral concepts and folk moral metaphysics were therefore centerstage in *TMP*, both the proper analysis of folk moral concepts and whether anything falls under them. My question was how I, as a representative member of the folk, think about the moral features of the world and whether the world has those features.

While this still seems to me to be a perfectly respectable philosophical project, I now think it was a mistake not to ask whether we are committed to thinking in terms of moral concepts, analyzed in the way I suggest. To be clear, I don't mean this to usher in a Sally Haslanger-style ameliorative project in metaethics (Haslanger 2012). The question isn't whether we should use such concepts, where the 'should' is moral or political or prudential. The question concerns the basic metaphysical structure of the world and the concepts we employ in grasping that basic structure. Are moral concepts, understood in the way I suggest, readily available to anyone with the ability to grasp that basic structure? I therefore now think of myself as only secondarily a representative member of the folk, and as primarily an old-fashioned metaphysician. Metaethics is properly thought of as a contribution to fundamental metaphysics, not just folk metaphysics.

One important consequence of this change is the attitude we should take towards metaethical disagreement. In *TMP* I argued that metaethicists who offer analyses of our moral concepts different from those I offer must have a false view about the concepts that they, who must also take themselves to be representative members of the folk, employ when they make moral judgements. But for the reasons just given, I no longer think it necessary to convict them of that kind of error. Perhaps folk moral thinking is more diverse than I imagined. Even if it is, what distinguishes moral concepts, understood in the way I suggest, is that they afford us a way of thinking about reality that is available to anyone who can grasp the basic metaphysical structure of the world. Of course, this difference between the concepts that we in fact use and those that are available to anyone who can grasp the basic metaphysical structure of the world would all but disappear if the basic metaphysical features of the world were Lewisian reference magnets for the words we use when we talk about features in the ballpark of those basic metaphysical features, as folk concepts would turn out to be constrained by fundamental metaphysics (Lewis 1984). But tempted though I am by this line of thought, I don't wish to commit myself to it. It suffices that moral features are metaphysically distinguished.

Another important consequence of this change concerns the normative ambition of metaethics. In *TMP* I thought that metaethics itself had few, if any,

first-order moral consequences. Though I assumed it was a conceptual truth that first-order moral views have a vaguely bounded content—more on this presently—I thought that this content was so vague as to be neutral between the best-known candidate moral theories like utilitarianism and deontology. But I now think that this too was a mistake. The basic metaphysical features of the world, I now think, tell in favour of a deontological moral theory, albeit one that has been consequentialized (Smith 2009). What are these basic metaphysical features that make such concepts available? They are that the world is spatio-temporally and causally ordered, that we figure in the actual world as a part, and that the distinctive feature of the part that is us is that we have and exercise the capacities to know what the world is like and satisfy our desires in it. In other words, reflective agents, understood in the way in which they are understood in the standard story of action, are themselves metaphysically distinguished, and this entails that all such agents are subject to deontological moral requirements.

The argument that the world has these basic metaphysical features traces back to Descartes, albeit with a non-Cartesian twist. We begin our philosophizing with a curious desire to figure out whether there is anything that we can know about the world based on reflection alone, in other words *a priori*, and we combine this desire with a belief that we can satisfy it by attending to that very question and seeing where our thinking leads us. This desire and means-end belief pair leads us to attend to that question and to have various thoughts. As a result of this process, we come to realize that we do know something *a priori*, namely, the contingent fact that we exist, thereby satisfying our desire, and we also know that attending to the question whether there is anything we can know on the basis of reflection alone and seeing where that leads us does indeed lead us to have knowledge of the world. Since the knowledge that we exist survives reflective scrutiny, and since it would be impossible for us to have such knowledge if we weren't agents in a spatio-temporal world that is causally ordered who perform mental acts like attending to questions and scrutinizing the answers, the fact that we have the knowledge that we exist and that we are agents thus presupposes that the world has the basic metaphysical features suggested earlier.

The argument for this conclusion is transcendental in nature. The conclusion is therefore strictly speaking disjunctive. Either we have no contentful thoughts at all, or the world in which we live isn't just knowable, but known by us to include ourselves as one of the things in it, that the parts that are us are agents, and that we are subject to deontological moral requirements. As we will see, beginning our metaethical thinking from this more basic metaphysical understanding of ourselves and the knowability of the world in which we live offers us an alternative route to the dispositional theory of value from that offered in *TMP*. This is good news, as I now think the route taken in *TMP* leads to a dead end.

2. Spelling out vs vindicating the folk conception of morality

In *TMP* the argument proceeds in two stages. At the first stage there is a spelling out, and at the second stage an attempt to vindicate, our folk conception of morality.

Along with the folk, I began the first stage by distinguishing between motivating and normative reasons for action. Motivating reasons are, I said, the psychological states that teleologically and causally explain our actions, where these are pairs of distinct intrinsic desires and means-end beliefs (this is Hume's theory of motivation), and normative reasons are considerations that justify our acting in the ways we act.¹ Following a long tradition in philosophy that goes back to Plato and Aristotle, I assumed that normative reasons for acting in a certain way are states of the world that we would bring about by so acting that are intrinsically desirable (or valuable, or good in the predicative sense—I didn't and still don't make distinctions between these, but I have now adopted the policy of talking exclusively in terms of desirability), and that the normative reasons themselves are just the natural intrinsic-desirability-making features of those states of the world. The nature of normative reasons is thus determined in part by what it is for a state of the world to be intrinsically desirable, and in part by what it is for agents

1 There has been a great deal of discussion about this distinction since *TMP* was published. Some argue that I was wrong to think of motivating and normative reasons as belonging to different categories: psychological states versus considerations (originally Dancy 1994–5, and most recently Howard and Schroeder 2024). In their view, motivating reasons should also be thought of as considerations, and more specifically as those considerations that move us when we deliberate and act. In “Backgrounding Desire” (1990), Philip Pettit and I had already argued that when (say) I desire to help those in need and believe that so-and-so is in need and can be helped by my ϕ -ing, and I ϕ as a result, there is a sense in which the desire and belief lie in the background of my ϕ -ing, and the putative fact that so-and-so is in need and can be helped by my ϕ -ing is in the foreground. For while the former causes my ϕ -ing, the latter is the consideration I attend to in deliberation when the former causes my ϕ -ing. I am happy to concede that there is a sense in which such foregrounded considerations are motivating, but I note that they are only motivating because of their relationship to the desires and beliefs in the background that explain my behaviour. I therefore think of the view that motivating reasons are considerations as an addendum to the view of them as psychological states, not an alternative to it. The view of them as psychological states is still explanatorily prior, and given that foregrounded considerations are only available to those agents who are capable of deliberation, it is still needed to explain the actions of non-deliberative agents. Some also claim that motivating and normative reasons are much more closely related than I say they are because all foregrounded considerations that are motivating must at least seem to us to be normative reasons, or be taken by us to be normative reasons, whereas I say there is no such constraint (Scanlon 1998). But my opponents' view on this matter is clearly false. Think about agents who are especially perverse. The considerations that move them may well be those that are known by them and seem to them to be considerations that *dyjustify* their actions, to use Michael Stocker's (1979) excellent term. Foregrounded motivating reasons need thus neither be nor seem to be normative reasons. (This footnote repeats some of what I say in a footnote to “The Revised Moral Problem” (Smith 2024b).)

who have those reasons to have the ability to bring those states about. The upshot is that, if our folk conception of morality is to be vindicated, then that vindication must in some way emerge from our folk understanding of intrinsic desirability.

In the second chapter of *TMP*, I had argued against cognitivist analyses of intrinsic desirability like Moore's, and Plato's too for that matter, that would require the intrinsic desirability of a state of affairs to be a non-natural property of that state of affairs on the grounds that such analyses cannot explain the supervenience of intrinsic desirability on natural features. I also argued that the only argument for a non-cognitivist analysis of intrinsic desirability judgements is a last-man-standing argument. If no plausible cognitivist analysis of intrinsic desirability can be squared with naturalism, then perhaps we should attempt to give a naturalistic non-cognitivist analysis of intrinsic desirability judgments instead. But the purpose of the rest of *TMP* was to prove by example that the antecedent of this conditional is false by coming up with a plausible naturalistic conception of intrinsic desirability.

My initial suggestion was that the states of the world that justify our actions—that is, the desirable states of affairs—should be understood as those natural states of affairs that would figure in the contents of the means-end beliefs and intrinsic desires that produce actions when those actions are immune to reasoned criticism. This initial suggestion gets refined in the fifth chapter of *TMP*, as it turns out that we can think of my counterpart who has reasoned-criticism-immune means-end beliefs and intrinsic desires as either an exemplar or an advisor, and that we should think of them as an advisor. But we can ignore this refinement for the time being (though see footnote 11). What's important is that, whether we think of my reasoned-criticism-immune counterpart as an exemplar or an advisor, the naturalistic credentials of intrinsic desirability turns on the naturalistic credentials of reasoned criticism.

At this point it is worth pausing to consider the relationship between *TMP*'s dispositional theory of desirability and a buck-passing theory of the kind proposed by Scanlon in *What We Owe To Each Other* (1998). Both theories pass the normative buck associated with desirability to a psychological state's having some feature. Both theories take that psychological state to be (inter alia) desire. For a Scanlon-style buck-passer, the feature of desire is *being one for which there is a reason*, whereas according to *TMP*'s dispositional theory, the feature of desire is *being immune to reasoned criticism*. So far the theories sound remarkably similar. The difference only emerges when we attend to what the two theories have to say about the nature of their preferred feature. The dispositional theory's feature will turn out to be a natural feature, whereas the Scanlon-style buck-passer's feature is a non-natural feature. What's at issue between the two theories is therefore the naturalistic credentials of reasoned criticism.

According to our folk understanding of reasoned criticism, I suggested in the fifth chapter of *TMP*, means-end beliefs are immune to reasoned criticism when they are part of an overall belief set that reflects neither ignorance nor error and hangs together in a maximally coherent way.² Nothing non-naturalistic is posited so long as we have naturalistic theories of truth, evidential support, and knowledge. Combinations of intrinsic desires and means-end beliefs are immune to at least one kind of reasoned criticism when they are part of a set where the strength of the resultant instrumental desires is proportional to the strengths of the various intrinsic desires and the confidence levels associated with the various means-end beliefs in the ways described by plausible structural theories of global instrumental rationality. Nothing non-naturalistic is posited here either.

Bernard Williams helpfully reminded us that agents are subject to reasoned criticism if they would intrinsically desire that the world is certain way if they vividly imagined what it would be like for it to be that way, but fail to have that intrinsic desire (Williams 1981). Elsewhere I have argued that Williams's suggestion is best interpreted as a coherence requirement linking the affective and motivational dispositions constitutive of intrinsic desire (Smith 2012). Coherence requires that those who are disposed to be pleased that *p* in virtue of *p*'s intrinsic features when they imagine that *p*, be disposed to act in certain ways when they believe that acting in those ways will bring *p* about. In *TMP* I also suggested that intrinsic desires are subject to reasoned criticism when they aren't suitably general, and when they aren't part of a maximally coherent and unified set.³ Again, nothing non-naturalistic is posited on formal and structural understandings of these ideas.

-
- 2 Note that being liable to reasoned criticism includes, but goes beyond, what the folk might call being irrational. What's required to avoid reasoned criticism is knowledge of means-to-ends, not just beliefs formed in the light of all the available evidence. This is as it should be given that in *TMP* the focus was on *objective* normative reasons for action, not *subjective* normative reasons. But once we know which of an agent's options would bring about intrinsically desirable states of the world—this is what he has an objective normative reason to do—we can figure out which ones would bring about states of the world that have expected desirability (think: expected utility) by plugging in the agent's expectations about means-to-ends and multiplying by how intrinsically desirable the outcome is, and we can also figure out which of his options would bring about states of the world that have responsibly-formed-expected desirability by plugging in those expectations about means-to-ends the agent would have if his expectations were formed in the light of all the available evidence and multiplying. In my view, if there is an unambiguous notion of what an agent has a subjective normative reason to do, then it corresponds to one or another of the latter.
 - 3 My thinking about this was influenced by what non-cognitivists must say about the rational pressures to which the desires we express when we make judgements about fundamental moral principles are subject. Since we are under rational pressure to revise these judgements via reflective equilibrium reasoning aimed at increasing their generality, coherence, and unity, non-cognitivists are committed to the view that the intrinsic desires—or intrinsic desires to intrinsically desire, or whatever they say

Towards the end of the fifth chapter, I argued that this does not exhaust our folk understanding of reasoned criticism. I suggested that means-end beliefs and intrinsic desires that are immune to reasoned criticism must be non-arbitrary, in some deep but yet-to-be-specified sense. In the case of belief, this sense is captured by the connection between reasoned criticism and truth, but it isn't obvious how to capture this idea in the case of intrinsic desire. The suggestion so far has been that intrinsic desires, which are neither true nor false, are subject to reasoned criticism when they are suitably general and part of a maximally unified and coherent set, and on the face of it this is consistent with their being arbitrary in the sense of their being idiosyncratic. People can and do intrinsically desire very different naturalistic states of affairs without their intrinsic desires failing to be suitably general, and without their intrinsic desire set seeming to display a failure of coherence or unity. If this kind of idiosyncrasy were reflected in our conception of intrinsic desirability, then that would immediately lead to the worrying kind of metaethical relativism that is inconsistent with the objectivity of morality.

To solve this problem, my suggestion in *TMP* was that for intrinsic desires to be non-arbitrary they must meet the further condition of being intrinsic desires that everyone who has means-end beliefs and intrinsic desires that are immune to reasoned criticism would converge upon. The attraction of this convergence requirement should be clear. Convergence in intrinsic desires with naturalistic content would rule out metaethical relativism without requiring us to postulate any non-natural features. But even if this is right and the concept of intrinsic desirability is the concept of having a feature that is the object of an intrinsic desire that everyone whose means-end beliefs and intrinsic desires are immune to reasoned criticism would converge upon, it doesn't follow that anything is in fact intrinsically desirable. At this point we move on to the second vindication stage of the argument.

States of affairs are intrinsically desirable, and hence people have normative reasons for action, I argued in *TMP*, just in case everyone whose means-end beliefs and intrinsic desires are immune to reasoned criticism would in fact converge in their intrinsic desires. A final vindication of morality therefore requires first, an argument for this claim about convergence in fact, and second, an account of the difference between the naturalistic contents of those converged upon intrinsic desires whose immunity to reasoned criticism entails that those naturalistic states of affairs have intrinsic moral desirability and those that have intrinsic non-moral desirability. As already mentioned, in *TMP* the latter account was a vaguely bounded naturalistic content constraint.

judgements of intrinsic desirability are expressions of—must themselves be subject to those same requirements of generality, coherence, and unity. While unsympathetic to their non-cognitivism, I think they're right about this. Some intrinsic desires are subject to such rational pressures. I am therefore unmoved by Derek Parfit's criticisms of this idea in Section 11 of *On What Matters* Volume One (2011). At best he shows that not all intrinsic desires are subject to such pressures.

Facts about intrinsic moral desirability were supposed to be constituted by reasoned-criticism-eluding intrinsic desires that people flourish (in some sense), or that they be treated with equal concern and respect (in some sense), and facts about intrinsic non-moral desirability were supposed to be constituted by converged upon intrinsic desires with completely different naturalistic contents, my thought being that these would all be variations on an intrinsic desire to satisfy whatever intrinsic desires one happens to have. With nothing non-naturalistic having been posited, so long as we have the argument for convergence in fact, the vindication would be complete.⁴

3. Why the proposed vindication of folk morality fails

Many critics of *TMP* focused their criticisms on the convergence requirement (see for example Horgan and Timmons 1996, Lenman 1999). One issue of interpretation that arose is whether the requirement was for a contingent or a necessary convergence. Since convergence was supposed to capture the non-arbitrary nature of intrinsic desires that are immune to reasoned criticism, and thereby rule out metaethical relativism, the answer had to be the latter.

Imagine that everyone actual intrinsically desired just one thing, that the world contains as much pleasure as possible, and hence that everyone actual converges in the intrinsic desires they would have if their intrinsic desires formed a maximally coherent and unified set. That would still be an arbitrary fact, in the sense of being a contingent fact, albeit one about everyone actual. Mere actual world convergence would therefore be inconsistent with the objectivity of morality, as in non-actual worlds where everyone intrinsically desired just one thing, but something different from pleasure, the moral facts would be different. Metaethical relativism would be the result. Accordingly, *TMP*'s bold suggestion was that what's required to avoid the arbitrariness of intrinsic desirability is a convergence in the intrinsic desires of every possible being whose intrinsic desires are immune to reasoned criticism.

This makes the critics concerns about convergence even more worrying. Imagining the requirement being met requires us to imagine there being something about reasoned criticism that explains why it is met. But what

4 In fact, there would still be work left to do, as though this would provide us with naturalistic accounts of moral and non-moral desirability, we wouldn't yet have naturalistic accounts of the standard deontic moral features: moral wrongness, moral permissibility, and moral obligation. In Chapter Six of *TMP* I did claim to provide an analysis of moral rightness in terms of being the object of a desire with a suitably restricted content that I would have if I were fully rational. But I wasn't explicit about how moral rightness relates to the standard deontic features, and until I was challenged in correspondence with Joshua Gert to say more about this some years later, I didn't have settled views. Another way in which my views have changed is that I have since abandoned the views I developed in that correspondence with Gert. For more on these changes, see the final section of this paper and Gert's contribution to this annual.

could that be? Given that intrinsic desires being suitably general and part of a maximally coherent and unified set is consistent with their being arbitrary—people can and do intrinsically desire very different naturalistic states of affairs without their intrinsic desires seeming to display a failure of generality or coherence or unity—that cannot be the explanation. Being suitably general and part of a fully informed set that displays maximal coherence and unity doesn't guarantee a necessary convergence either, given that the information is restricted to means-ends information. So when we imagine ideal reasoners converging in their intrinsic desires, what do we imagine the explanation of that convergence to be? In *TMP* I proposed the following answer.

In the second and fifth chapters of *TMP* I had argued that there is a wide-scope local coherence requirement linking beliefs about desirability and desires. If we imagine non-ideal reasoners attempting to get their psychology into a state of reflective equilibrium—that is, attempting to simulate ideal reasoners—then as we imagine them changing their beliefs about which features of states of affairs are intrinsically desirable under the pressure of argument, we also imagine there being corresponding changes in the intrinsic desires they have.⁵ For intrinsic desires to elude reasoned criticism, they must therefore be suitably general and part of a maximally coherent and unified set of beliefs and desires that ideal reasoners converge upon in part because of the local coherence of their intrinsic desires with their beliefs about intrinsic desirability, and in part because those beliefs about intrinsic desirability themselves elude reasoned criticism.

To see what I had in mind here, think about non-ideal reasoners engaging in first-order reflective equilibrium arguments of the kind we find in normative ethics, and for these purposes imagine that their target question about intrinsic desirability is simply which intrinsic desires ideal reasoners would converge on. If they exist in circumstances that augur in favour of reasoned-criticism-eluding beliefs and desires—and my claim in *TMP* was that, as the influence of religious thinking declines and we become better

5 Though *TMP* predates Judith Jarvis Thomson's *Normativity* (2008), this idea is clearly in the same ballpark as a much deeper insight of hers. Thomson suggests that a reason for being in a psychological state with a correctness condition is a consideration that supports the truth of that psychological state's correctness condition. For example, a reason for believing *p* (belief is a psychological state with a correctness condition) is a consideration that supports the truth of *p* (that belief's correctness condition). But now consider the intrinsic desire that *p*. This too is a psychological state with a correctness condition, that correctness condition being the intrinsic desirability of *p*. Reasons for intrinsically desiring that *p* are those considerations that support the truth of the claim that *p* is intrinsically desirable. If Thomson is right, then it should come as no surprise that as non-ideal reasoners change their beliefs about what's intrinsically desirable, under pressure of what they take to be reasons for these changes, their attempt to achieve a reflective equilibrium in their overall psychology will usher in corresponding changes in their intrinsic desires. That follows from the fact that both belief and intrinsic desire are sensitive to reasons for, in the sense of considerations that support the truth of, the truth of their very different correctness conditions (see also Smith 2012).

at critical thinking, more and more of us will exist in such circumstances—then over time we could expect these arguments to get more refined, and the reasoners themselves to become more sensitive to them, losing their conviction in intrinsic desirability claims that gain no support from them. Their beliefs about intrinsic desirability, and their intrinsic desires too, would become part of an increasingly informed and coherent and unified set. As such reasoners get more ideal, I suggested, we could expect there to be a convergence in reasoner's beliefs about intrinsic desirability, and hence in their intrinsic desires. At the limit, ideal reasoners would all have the same beliefs about intrinsic desirability and the same intrinsic desires, or so I suggested.

The guiding idea in all of this was that folk morality is committed to the truth of moral rationalism: that is, to the truth of the claim that moral facts, if there are any, entail categorical facts about agents' normative reasons, reasons that are not conditional on the agents who have such reasons having certain intrinsic desires rather than others. It was this background commitment of folk morality to moral rationalism that led me to think that facts about intrinsic desirability are not facts about the arbitrary intrinsic desires that agents actually have, but are instead facts about those that they would have if their intrinsic desires eluded reasoned criticism, where the categoricity of such normative reasons is captured by the required convergence in agents' reasoned-criticism-eluding intrinsic desires.⁶ This commitment of folk morality to the truth of moral rationalism, I thought when I wrote *TMP*, supports the crucial premise in the line of reasoning outlined above, the premise that ideal reasoners would converge in their moral beliefs. This is because moral rationalism itself implies that if there are any pure moral facts—that is, moral facts that don't imply contingent a posteriori non-moral facts—then those facts aren't just, if known, known a priori, but are known to ideal reasoners. Moral facts that remain hidden even to ideal reasoners would, after all, be unable to play the role that moral facts are supposed by moral rationalism to play, the role of making it possible for everyone to get along by freely doing what they have categorical normative reason to do.⁷

However, this line of reasoning in support of the claim that ideal reasoners converge in their intrinsic desires faces a serious problem (Kelly and McGrath 2010). The rational credibility of the first-order moral conclusions of reflective equilibrium arguments given in normative ethics depends on

6 Note in passing that this is to move from a plausible claim about moral intrinsic desirability to a claim about intrinsic desirability in general. I will have more to say about this move in the final section's discussion of the second payoff.

7 So while Sarah McGrath (2013/2010) is right that versions of moral realism without convergence are available, those who are realists because they are moral rationalists should reject such versions. They should instead accept the version of moral realism implied by the argument for moral rationalism given in the fourth section of this paper. See the final payoff in the final section.

the rational credibility of their premises. This is why reflective equilibrium arguments have the well-known garbage-in, garbage-out problem. The line of reasoning in *TMP* for the conclusion that, if there are moral facts, ideal reasoners will converge in their first-order judgements of moral intrinsic desirability, and hence in their intrinsic desires, therefore requires us to give an account of which reasoners reach their conclusions based on premises that themselves have rational credibility—these will be the ideal reasoners—and which do not. Importantly, however, this account cannot on pain of circularity presuppose that, if there are moral facts, ideal reasoners will converge in their intrinsic desires—and hence cannot on pain of circularity presuppose that ideal reasoners will converge upon certain intrinsic desires rather than others—as the optimistic line of argument is supposed to establish that there are moral facts because ideal reasoners converge in their intrinsic desires.

Consider two imaginary figures in contemporary normative ethics engaging in this process: Shinger and Shcanlon. Both are excellent philosophers, so each of them does their best to get their psychology into a state of reflective equilibrium, incorporating an evaluation of the other's premises into their own psychology, and coming up with grounds for rejecting those premises satisfactory to themselves. Shinger ends up with utilitarian intrinsic desirability beliefs and desires and Shcanlon ends up with deontological-of-the-contractualist-kind intrinsic desirability beliefs and desires, so neither of them thinks of the other as their epistemic peer.⁸ Is either of them an ideal reasoner? For instance, do either of them make a mistake when they incorporate an evaluation of the other's premises into their own psychology, and so argue on the basis of premises that lack rational credibility? The problem is that the obvious answer to give to this question cannot be given. We cannot say that the premises with rational credibility are those that would enable them, by getting their psychology into a state of reflective equilibrium, to have knowledge of what's intrinsically desirable, if anything is, because that obvious answer requires an account what the intrinsic desirability facts are, if there are any, and the reason we need an account of rational credibility is in order to establish precisely that.

Though I was aware of this problem and tried to solve it, the solution I proposed was inadequate. I said that the reasoners who do not reach their conclusions based on premises with rational credibility are those who are guilty of either group-think or intellectual arrogance (*TMP* pp. 188–9, 194–6). But I failed to show that we could say what either of these are without presupposing which first-order judgements of intrinsic desirability, and hence which intrinsic desires, ideal reasoners would converge upon. The upshot is that the attempt to vindicate the existence of intrinsically desirable states of affairs at the end of *TMP* is a failure. Worse still, if *TMP* succeeds in showing that folk morality is committed to the truth of moral rationalism,

8 Here I assume that contractualism can be consequentialized (Smith 2009, Portmore 2011).

the fact that moral rationalism requires a necessary convergence in the beliefs about intrinsic desirability and intrinsic desires of ideal reasoners, if that is indeed a fact, and the difficulties that emerge when we try to argue for such a necessary convergence in fact, suggests that moral rationalism is false. So if we restrict ourselves to the arguments given in *TMP*, it is no surprise that some think it shows that we should be error theorists about moral intrinsic desirability (see Joyce 2001).

4. Fundamental metaphysics to the rescue

As I said at the outset, the biggest change in my thinking since writing *TMP* concerns the metaphysical ambition of metaethics. In the past I thought the only way to argue for first-order moral claims was via reflective equilibrium arguments of the kind given in normative ethics, the kind we imagine Shinger and Shcanlon rehearsing to themselves as they attempt to get their own psychologies into a state of reflective equilibrium. But I no longer think this.

Shinger and Shcanlon give arguments from above, in the sense that they begin from the assumption that they have all sorts of justified moral beliefs, and that the contents of those in which they have the greatest confidence not only can reasonably, but must, figure as premises in the reflective equilibrium arguments they give for first-order moral conclusions. When I wrote *TMP*, I thought that arguments from below for first-order moral claims—that is, arguments that do not proceed from that assumption—were doomed to fail. Such arguments are famously associated with Kant, but you also find Humean versions in the work of David Gauthier (1987) and Gilbert Harman (1975).⁹ But while the arguments from below these theorists provide do indeed fail, I no longer think such arguments are bound to fail, because I now see how first-order moral claims could follow from the basic metaphysical truths outlined earlier. This is good news, as it would explain why not just humans agents, but all possible reflective agents, have the same moral obligations and can be held responsible for acting morally wrongly when they have no excuse. Since I've

9 Kant argues that the only actions we can consistently will to be a universal law of nature are those that treat humanity always as an end, never merely as a means. Unfortunately, though I believe the conclusion, or something close to it, I don't understand the argument Kant gives well enough to evaluate it. Gauthier's and Harman's arguments are much easier to understand, but they both entail that the class of agents who have moral obligations is implausibly restricted. Gauthier thinks that only those with a translucent psychology have moral obligations, which entails the false claim that many normal adult humans do not have moral obligations. Harman thinks that only those with whom I share cooperative intentions have moral obligations, and that the content of our obligations is fixed by the contents of our shared intentions. This entails the false claim that all those with whom I do not share cooperative intentions—think of adult humans with radically different moral beliefs from mine, and powerful selfish adult humans who have no need to share cooperative intentions with me—have no moral obligations.

published variations on this argument elsewhere (for example Smith 2020), I will not repeat it fully here. I will, however, provide sufficient detail to make it clear how my current views differ from those I held when I wrote *TMP*.

Let's return to the task of spelling out a naturalistic conception of desirability. Given the role of intrinsic desirability in providing us with justifications for acting, and the role of means-end beliefs and intrinsic desires in producing action, my suggestion in *TMP* was that the states of the world that justify our actions—that is, the desirable states of affairs—should be understood to a first approximation to be those natural states of affairs that would figure in the contents of the means-end beliefs and intrinsic desires that produce actions when those actions are immune to reasoned criticism. The strategy in *TMP* was to spell out our folk conception of reasoned criticism, noting *en passant* that it didn't require us to posit anything non-natural. There is, however, a different route to this same conclusion, a route that begins from the fact that we know a priori the contingent fact that we exist and that we are agents. This route not only vindicates the naturalistic credentials of reasoned criticism that remains central to understanding the nature of desirability, but also entails first-order conclusions about what's intrinsically desirable both morally and non-morally, and thus what we have normative reasons to do. It therefore solves the problem with reflective equilibrium reasoning that arose in the case of Shinger and Shcanlon.

The crucial initial point is that the kind *agent* is what Judith Jarvis Thomson calls a *goodness-fixing kind*, where this is goodness in the attributive sense, not the predicative sense. We are therefore firmly in argument from below territory, as no moral assumptions are being made. Goodness-fixing kinds are kinds that allow us to order members of the kind from best to worst according to a standard of assessment internal to the kind itself. For example, the kind *parasite* picks out a kind of living thing that lives in and off a host, which is another living thing, getting their nourishment from the host at the host's expense. Individual parasites can therefore be ordered from best to worst according to how successful they are at doing that. The kind *agent* is also a goodness fixing kind, as agents are a kind with the capacity to form beliefs about means-to-ends that combine with intrinsic desires to produce behaviour aimed at the satisfaction of those intrinsic desires, and they too can therefore be ordered from best to worst according to how successful they are at doing this. But what exactly is the standard internal to this ordering? The distinction between motivating and normative reasons once again comes to fore.

We know that the class of agents includes beings much simpler than humans. Agents are simply beings whose behaviour is sufficiently complex to warrant motivating reasons—that is, intrinsic desires and means-end beliefs—as their explanation. Since innate drives that connect up with perceptual representations of a being's environment and motor system are presumably intrinsic desire and means-end belief in their most primitive form, the class of

agents includes, as well as adult humans and beings in other possible worlds of a different species who are like adult humans, infant humans, kangaroos, magpies, great white sharks, and perhaps even some insects. Normative reasons are, however, different. The possession of normative reasons requires at least the capacity to believe that a failure to do what one has a normative reason to do, absent an excuse or a justification, makes one liable to reasoned criticism.¹⁰ Since infant humans, kangaroos, magpies, great white sharks, and insects clearly lack the conceptual sophistication required to have such a belief, they do not have normative reasons, though they do have motivating reasons. A sensitivity to normative reasons is, however, the main driver in the ordering of agents from best to worst.

Ideal agents, those at the top of the ordering, are those who robustly have and exercise maximal capacities to do what they have normative reason to do. In other words, for any circumstance in which there is a normative reason available to do something, including circumstances that they aren't in but could be in, they have that reason; their desires and means-end beliefs perfectly align with that reason; and these facts about them are maximally non-accidental. There is therefore a sense in which such agents are prepared for every contingency. Less ideal agents are those for whom there is less of an alignment along at least one of these dimensions. They have fewer skills, they are less knowledgeable about normative reasons or means-to-ends, they are less instrumentally rational, or, though they aren't deficient in any of these respects, their not being so is to some extent fluke. As we go further down the ordering, we find agents who lack normative reasons for action altogether, though they still resemble such beings insofar as they have intrinsic desires and means-end beliefs and are instrumentally rational to some degree. As we approach the bottom of the ordering, though their behaviour is still sufficiently complex to warrant explanation in terms of intrinsic desires and means-end beliefs, it is only just so.

Note that the various dimensions of this ordering provide us with non-question-begging tests of both the plausibility of substantive theories about what agents have normative reason to do and their naturalistic credentials, and that they do so without begging any questions about the nature of normative reasons. Focus on ideal agents, the worlds in which they live, and the capacities for knowledge-acquisition and desire-realization that different substantive theories of what agents have normative reason to do tell us they have and exercise. These theories start from a shared concept of an ideal agent as someone who robustly has and exercises their capacity to do what they have normative reason to do, but go on to fill in the details in terms of different conceptions. As between any two such theories, the more plausible theory is that with the better mix of two features. The first is parsimony. One theory is

10 I suspect that the normative foundation of the M'naughton Rule in the law is the fact that this capacity is required for an agent to have normative reasons.

pro tem better than another if it explains at least as much as the other but with less. Naturalism is therefore the default, given that we are committed to the existence of natural features willy nilly, but non-naturalism is not ruled out. The second is robustness. One theory is pro tem better than another if ideal agents, understood in terms of that theory's conception, possess and exercise their agential capacities more robustly than do ideal agents understood in terms of the other theory's conception. The best conception of an ideal agent is that associated with the best theory of normative reasons.

Consider now the Humean's theory. What is the Humean's conception of an ideal agent, what does that conception tell us about normative reasons, and what does it tell us about the nature and naturalistic credentials of reasoned criticism? For each non-ideal agent at a time, the Humean holds that there is an ideal counterpart of that agent whose intrinsic desires and means-end beliefs are not identical to the non-ideal agent's, but are a function from them. As before, their intrinsic desires are those that the non-ideal agent at that time would have if his intrinsic desires were immune to reasoned criticism. For the Humean, all there is to a non-ideal agent's intrinsic desires at a time being immune from reasoned criticism is their being those that that agent's counterpart would have if his intrinsic desires were suitably general and part of a maximally unified and coherent set. And as before, the ideal agent's means-end beliefs are those the non-ideal agent would have if he were neither ignorant nor in error about means-to-ends. In other words, if he had total means-end knowledge. We can think of the non-ideal agent's ideal counterpart's intrinsic desires at that time about the states of the non-ideal agent's world at any time as fixing the facts about what's intrinsically desirable at any time in the non-ideal agent's world, relative to that agent at that time, and the instrumental desires the ideal agent has at that time, given his global instrumental rationality and his knowledge of the means available to the non-ideal agent to satisfy his intrinsic desires at that time, as fixing the facts about which of the non-ideal agent's options are instrumentally desirable relative to that agent at that time, and thus which options he has normative reasons to pursue (from here-on I will take these cumbersome relativizations of desirability as read).

Note that this more metaphysically driven account of the relationships between desirability and reasoned criticism, and normative reasons and desirability, is the same as the folk conception given in *TMP* before the addition of the requirement of non-arbitrariness. Its naturalistic credentials are therefore secure. But the focus on ideal agents, and more especially the robustness of ideal agent's doing what they have most normative reason to do, is helpful because we can already see that the account of reasoned criticism is incomplete. Ideal agents have the capacity for self-control: that is, the capacity, when they believe themselves to have most normative reason to act in a certain way, to get themselves to desire most instrumentally to act, and

then act, in that way. Imagine two otherwise identical agents with whatever natures we imagine ideal agents to have minus the capacity for self-control, one of whom has that capacity and the other of whom lacks it. In the nearest non-ideal worlds to the ideal world in which they don't desire most strongly to do what they have most normative reason to do, the one who possesses the capacity for self-control exercises it and gets himself to act in that way anyway, whereas the one who lacks the capacity for self-control doesn't. The one with that capacity therefore does what he has most normative reason to do more robustly than the one who lacks it.

The upshot is that, whatever the correct conception of an ideal agent, an agent is apt for reasoned criticism if he either lacks self-control, or he has it but his actions can be explained by his failure to exercise it. But does the addition of the capacity for self-control threaten the naturalistic credentials of our understanding of an ideal agent? For example, does it require agent-causation, or an exercise of libertarian free will? My own view is that it is an empirical question what the mechanisms of self-control are, and that these mechanisms may differ from person to person and within a person from time to time. But let me give an example of a naturalistic disposition agents could manifest that even Humean naturalists should admit would qualify as their exercise of the capacity for self-control. Imagine an agent who meets the description of the Humean ideal agent, prior to the addition of the capacity for self-control, but who additionally has the disposition, whenever he believes himself to have most normative reason to act in a certain way but does not instrumentally desire most to act in that way because of his global instrumental irrationality, to pinch himself, and imagine further that the resulting pain he feels causes him to most instrumentally desire to do what he believes he has most normative reason to do, and that he in turn does what he has most normative reason to do.

There are in fact two quite different things a Humean might imagine when imagining this. He might imagine the agent having an intrinsic desire to be rational and a belief that he can make himself rational by pinching himself with respect to which he is not instrumentally irrational, and hence that his pinching himself is itself an action, or he might imagine the pinching not to be caused by an intrinsic desire and means-end belief, but instead to be a reflex triggered by his belief that he is being instrumentally irrational. Whichever of these the Humean imagines, the manifestation of such a disposition would seem to qualify as an exercise of the capacity for self-control. Given that what's imagined doesn't include agent-causation, or the exercise of libertarian free will, or any other non-natural feature for that matter, it seems that the possession and exercise of self-control can therefore be understood in wholly naturalistic terms. The naturalistic credentials of the Humean conception of normative reasons in terms of the desires of an ideal

agent, amended to include that agent's having the capacity for self-control, are therefore secure.¹¹

This brings us to the all-important question. Is an ideal agent's doing what he has most normative reason to do a robust fact about him, according to the amended Humean conception? Given the intra- and inter-agential obstacles such an agent faces, the answer is that though it is somewhat robust, it is also somewhat fragile. The ideal agent's self-control overcomes a synchronic intra-agential obstacle to his doing what he has most normative reason to do at a time, namely his own instrumental irrationality. But the potential for conflict between the intrinsic desires for the future he has at earlier times, and those he has for the future at those future times, remains a significant diachronic intra-agential obstacle to his doing what he has most normative reason to do over time.¹² To secure the fact that he does indeed do what he has most normative reason to do over time, the amended Humean conception must therefore stipulate that in the ideal world that potential isn't realized because the intrinsic desires ideal agents have at different times have contents that in some way harmonize with each other. Because his intrinsic desires harmonize, in doing what he has most normative reason to do at an earlier time, the ideal agent doesn't destroy his future capacity to do what he has most normative reason to do, or stand idly by while it diminishes, and nor does he interfere with his future exercise of that capacity either.

The obvious way to secure such harmony within an agent would be by stipulating that he has two two additional suitably strong intrinsic desires.

11 The addition of self-control to our understanding of an ideal agent makes it clear why we must interpret normative reasons in terms of the desires of an advisor, not an exemplar. Whether the non-ideal agent has a normative reason to do what his ideal counterpart does—this is what the exemplar interpretation tells us—depends on the costs associated with his doing that via an exercise of self-control. To make things simple, imagine that the non-ideal agent desires just one thing, to get as much pleasure and as little pain as possible, and that the action that is globally instrumentally rational in his circumstances is forgoing the additional short-term pleasure associated with eating a chocolate and eating an apple instead, thereby getting more pleasure in the long-term. Since the ideal agent is globally instrumentally rational, this is what he most desires to do and does, and is therefore what he has most normative reason to do and does in his world. However, let's suppose that the salience of the chocolate makes the non-ideal agent instrumentally irrational, and as a result he desires more to eat the chocolate. Does the non-ideal agent have most normative reason to eat the apple, as the exemplar interpretation implies, given that the mechanism of self-control requires him to experience pain? The correct answer is that given by the advisor interpretation. It all depends on whether his ideal counterpart, who desires the non-ideal agent to experience as much pleasure and as little pain as possible in the long-term, desires most that the non-ideal agent eats the apple, given the additional pain that he would have to experience in order to do so. If the pain is too great, then his ideal counterpart will not desire most that he eats the apple, so the non-ideal agent will not have most normative reason to do so.

12 To my knowledge, this observation was first made by Thomas Nagel (1970), though not in the service of making the present point, which is that the ideality of ideal agents, on the amended Humean understanding, is an implausibly fragile feature of them.

The first is an intrinsic desire that he has and maintains the capacities for desire-realization and the acquisition of knowledge of means in the future, and the second is an intrinsic desire not to interfere with either his future acquisition of knowledge of means or his realization of future intrinsic desires, the latter on condition that their realization doesn't require that he interferes with intrinsic desires or knowledge-acquisition in the further future. But the Humean cannot, on pain of giving up on his Humeanism, say that a failure to have such intrinsic desires makes an agent liable to reasoned criticism. For remember, all there is to a non-ideal agent's intrinsic desires at a time being immune from reasoned criticism for the Humean is their being those that that agent's counterpart would have if his intrinsic desires were suitably general and part of a maximally unified and coherent set. According to the amended Humean conception, an agent who lacks the intrinsic desires an ideal agent must have if he is to do what he has most normative reason to do over time is therefore, paradoxically, immune from reasoned criticism. Worse still, while this stipulation about the intrinsic desires the ideal agent must have guarantees that he does what he has most normative reason to do at both earlier and future times in the ideal world, it also guarantees that in those nearby non-ideal worlds to the ideal world where his intrinsic desires over time don't harmonize, his doing what he has most normative reason to do at the earlier time may require either that he interferes with his capacity to do what he has most normative reason to do at the future time, or destroys it, and hence that he does not do what he has most normative reason to do at that later time. His doing what he has most normative reason to do is therefore, to this extent, fragile.

Moreover, the stipulation does nothing to address the significant synchronic inter-agential obstacles to an agent's doing what he has most normative reason to do. As before, the amended Humean conception must stipulate that (say) whenever the ideal agent acts in an ideal world with other agents in it, one of the following is true of those other agents. Either they have intrinsic desires that harmonize in some way with his, or though they don't harmonize with his he is more powerful than they are and so can dominate them, or though they don't harmonize with his and he is not more powerful than them, they are causally isolated from him. But again, in those nearby non-ideal worlds to the ideal world in which these conditions aren't met, other agents do interfere with his exercise of his capacity to do so, and perhaps even destroy it altogether, so he doesn't do what he has most normative reason to do. The amended Humean conception of normative reasons in terms of the desires of an ideal agent is therefore one according to which the ideal agent's doing what he has most normative reason to do is, in these respects too, fragile.

The fragility of an ideal agent's doing what he has most normative reason to do, given the amended Humean conception of an ideal agent, suggests an argument for an anti-Humean conception of normative reasons that should

be convincing even to Humeans. Imagine that the anti-Humean comes up with a conception of an ideal agent which is just as naturalistic as the amended Humean conception, which makes use of an account of reasoned criticism that is as credible as the amended Humean's, and which entails that the ideal agent's doing what he has most normative reason to do is more robust than it is according to the amended Humean conception. It should be agreed by all concerned that this would provide a strong reason to prefer that anti-Humean conception of normative reasons to the amended Humean conception. It is perhaps already clear what the details of such an anti-Humean conception should be.

First and foremost, the anti-Humean should propose that an agent's lacking those intrinsic desires that they must have in order to robustly do what they have most normative reason to do makes them liable to reasoned criticism, and hence non-ideal. He should then propose that for each non-ideal agent at a time, there is an ideal counterpart of that agent whose intrinsic desires and means-end beliefs are not identical to the non-ideal agent's, but are those that the non-ideal agent at that time would have if his intrinsic desires and means-end beliefs were immune to other standards of reasoned criticism. And finally, he should propose that ideal agents live in a world in which their world-mates are themselves ideal. In other words, what makes ideal agents ideal is in part their living in a world with other ideal agents.

In more detail, the anti-Humean should agree with the amended Humean conception that being immune to reasoned criticism requires having those intrinsic desires that the non-ideal agent would have if his intrinsic desires were suitably general and part of a maximally unified and coherent set, and also having, as well as exercising when necessary, the capacity for self-control. But he should part company with the amended Humean conception's account of the intrinsic desires that are required for an ideal agent to robustly do what he has most normative reason to do. He should instead say that this requires having both a suitably strong intrinsic desire that all agents have the capacities for desire-realization and the acquisition of knowledge of means (for short, let's call this an intrinsic desire that everyone be helped) and a suitably strong intrinsic desire not to interfere with anyone's acquisition of knowledge of means or the realization of their intrinsic desires, on condition that the realization of those intrinsic desires doesn't require that other intrinsic desires be interfered with (for short, let's call this an intrinsic desire not to interfere). The ideal agent's task is then to balance these various intrinsic desires against each other, given their strengths, when they come into conflict, not only in the circumstances in which they find themselves in their own ideal world, but also in every other circumstance in which they might find themselves, including the circumstances in which the non-ideal agent finds himself in his world. Note that nothing about reasoned criticism, so understood, requires us to postulate anything non-natural. The amended Humean and anti-Humean conceptions of normative reasons are therefore on a par with regards to parsimony.

When we imagine the ideal agent balancing his various intrinsic desires against each other, we can think of this as his figuring out the relative weights that different intrinsic desirability-making features have vis a vis each other. For at a time, the non-ideal agent's ideal counterpart's intrinsic desires about the states of the non-ideal agent's world at that time fix the facts about what's intrinsically desirable in the non-ideal agent's world, relative to that agent at that time, and the instrumental desires the ideal agent has at that time, given his global instrumental rationality and his knowledge of the means available to the non-ideal agent to satisfy his intrinsic desires at that time, fix the facts about which of the non-ideal agent's options are instrumentally desirable relative to that agent at that time, and thus which options he has normative reasons to pursue. But note that with certain intrinsic desirability-making features these relativizations become less relevant. Agents in general having the capacities to realize their intrinsic desires and acquire means-ends knowledge is intrinsically desirable relative to every agent at every time, and every agent is such that their not interfering with anyone's exercise of these capacities at some time is intrinsically desirable relative to them at that time.

To summarize, though both are thoroughly naturalistic, there are three main differences between the amended Humean conception of normative reasons and the anti-Humean conception. The first difference, as I've already said, is that according to the anti-Humean conception, an agent's lacking those intrinsic desires he must have in order to robustly do what he has most normative reason to do makes him liable to reasoned criticism. In other words, the anti-Humean rejects the amended Humean's paradoxical view that though certain intrinsic desires are required for an ideal agent to robustly do what he most normative reason to do, he would not be liable to reasoned criticism if he lacked such desires. The second difference is that the intrinsic desires the anti-Humean thinks agents must have, in order to robustly do what they have most normative reason to do, are the intrinsic desires that everyone be helped and that he not interfere with anyone. They are not intrinsic desires that they be helped and that they do not interfere with just themselves. The third difference is that the ideal agent's world-mates are themselves ideal. They aren't just agents whose intrinsic desires harmonize with his, or agents he can dominate, or agents from whom he is causally isolated. The reasons for these differences are just what you would expect. These differences guarantee that an ideal agent's doing what he has most normative reason to do, according to the anti-Humean conception, is a much more robust fact about him than it is according to the amended Humean conception.

Consider the anti-Humean ideal agent in his ideal world. Is his doing what he has most normative reason to do over time vulnerable to his lacking the intrinsic desires that he must have if he is to robustly do what he has most normative reason to do? Not to the same extent as it is on the amended

Humean conception. For in the nearest non-ideal world to the ideal world in which he (say) lacks the intrinsic desire not to interfere with anyone, and hence lacks the intrinsic desire not to interfere with himself later, he exercises self-control and gets himself to desire to do, and do, what he has most normative reason to do, which is not to interfere with himself later. Note that what's doing all of the work here is the first and second of the three main differences just mentioned, combined with the view of the relationship between self-control and reasoned criticism that both the amended Humean and anti-Humean conceptions agree on. An agent's lacking the intrinsic desires he must have if he is to robustly do what he has most normative reason to do—that is, in this case, his lacking the intrinsic desire not to interfere with anyone—makes him liable to reasoned criticism, reasoned criticism that he can avoid by having and exercising self-control.

Is the ideal agent's robustly doing what he has most normative reason to do at a time vulnerable to other agent's lacking the intrinsic desires that they must have if he is robustly to do what he has most normative reason to do at that time? Not to the same extent as it is on the amended Humean conception. For in the nearest non-ideal world to the ideal world in which other agents lack the intrinsic desire not to interfere with anyone, and hence lack the intrinsic desire not to interfere with him at some time, they exercise self-control and get themselves to desire to do, and do, what they have most normative reason to do, which is not to interfere with him at that time. Note that what's doing all of the work this time is all three of the differences just mentioned, combined with the view of the relationship between self-control and reasoned criticism that both the amended Humean and anti-Humean conceptions agree on. Other agents lacking the intrinsic desires that they must have if a particular agent is to robustly do what he has most normative reason to do—that is, other agents lacking *inter alia* the intrinsic desire not to interfere with anyone—makes them liable to reasoned criticism, reasoned criticism that they can avoid by having and exercising self-control.

We saw earlier that a theory of normative reasons for action in terms of the desires of an ideal agent, given some conception of an ideal agent, is *pro tem* better than another in terms of some other conception if it explains at least as much as the other but with less, and if ideal agents, understood in terms of that conception, possess and exercise their agential capacities more robustly than do ideal agents understood in terms of the other conception. What we've just seen is that the anti-Humean conception is as good as the amended Humean conception on the first measure (both are naturalistic) and better on the second (ideal agents do what they have most normative reason to do more robustly). Of course, this doesn't show that the anti-Humean conception of an ideal agent is the best such conception, but let's leave the argument for that conclusion for another day and proceed on the assumption that it is for the remainder.

5. Five payoffs

The first payoff of the argument from below just given for the anti-Humean conception of normative reasons for action is that it establishes the truth of a necessary convergence thesis, albeit a more limited one than that argued for in *TMP*. All possible ideal agents have intrinsic desires that everyone be helped and that they themselves do not interfere because their having such intrinsic desires is part of what makes them ideal. It therefore follows that the normative reasons for action that these intrinsic desires ground are themselves reasons for action that all ideal and non-ideal agents have independently of the intrinsic desires that they may happen to have. They are therefore categorical reasons for action.

A second payoff of the argument from below is that it provides us with a better way of picking out the moral normative reasons for action from the non-moral ones. As we have just seen, moral normative reasons for action are categorical and non-moral normative reasons for action are hypothetical. Though it turns out to be true that these categorical normative reasons do indeed have content of the kind I proposed moral reasons have, for there is a sense in which they are just reasons to ensure both that everyone flourishes (the reason to ensure that everyone has agential capacities to exercise) and that we treat each other with respect (the reason not to interfere with anyone's exercise of their agential capacities), their distinctive content is a consequence of their categoricity. Suitably strong categorical normative reasons for action with such contents are what ensure that ideal agents all robustly do what they have most normative reason to do.

But what about the hypothetical reasons? Non-moral normative reasons for action are hypothetical because they are grounded in intrinsic desires that ideal agents have, but which their lacking would not count against their being ideal. They are therefore arbitrary in the sense I was so worried about when I wrote *TMP*. Is this something that we should be worried about? It is not (see again footnote 6). The arbitrariness of non-moral normative reasons is a reflection of the fact that it is up to each of us to decide how we want to live our lives based on whatever intrinsic desires we happen to find ourselves with. Even in the ideal world, different ideal agents may choose to lead their lives in very different ways from each other, and may choose to lead their lives in very different ways at earlier and later times. The arbitrariness of their intrinsic desires is what makes this kind of choice possible. I was therefore wrong to insist on the non-arbitrariness of non-moral normative reasons for action in *TMP*.

A third payoff of the argument from below just given is connected to what might at first seem to be a problem for the conclusion of that argument (compare Bukoski 2016). Crudely put, what the argument shows is that we all have normative reasons to do three things—to help, not to interfere, and apart

from that to do whatever want—that must be weighed against each other, but the argument leaves the content of these reasons and their weights extremely vague. The only guide it provides to help us think more precisely about these contents and weights is an image of the possible world in which everyone is ideal and so does what they have most normative reason to do at every moment they exist, and therefore a sense of what's required for a similarity between the ideal world and those non-ideal worlds in which all, nearly all, many, and so on, non-ideal agents do what they have most normative reason to do. Is this worrying? I do not think so.

There are two ways we might address the vagueness of the argument's conclusion. One is to stop doing philosophy and begin the practical task of getting everyone to agree on social rules that give these reasons more determinate content and weight, albeit as constrained by the three reasons with their vague contents and weights (Smith 2021). The other is to do more philosophy, and in particular, to do more philosophy of the kind that Shinger and Shcanlon do by engaging in reflective equilibrium reasoning to figure out whether there are reasons to prefer giving these reasons more determinate content and weight in certain ways rather than others. At this point we return to the garbage-in, garbage-out worries we had about reflective equilibrium reasoning for first-order normative ethical conclusions that plagued the argument for convergence in fact at the end of *TMP*. But this time we can ensure that we do not put garbage in when we begin our reasoning. What should the inputs to this reflective equilibrium reasoning be? The answer is that they should not be those first-order moral claims in which we have greatest antecedent confidence, but should instead be the conclusions of the argument from below.

In other words, we should engage in reflective equilibrium reasoning with the vaguely bounded moral reasons to help but not interfere, and the non-moral reason to do whatever we want to do, together with the image of the possible world in which everyone is ideal and so does what they have most normative reason to do at every moment they exist, as inputs, and the similarities of that world to those non-ideal worlds in which agents do much the same thing, and see whether certain ways of giving these reasons more determinate contents and weights turn out to be more compelling than others. But when we engage in such reasoning we must be mindful of the fact that, even if we were to make considerable progress, it would be unlikely that we would remove all of the need for social rules to give the reasons to help, not interfere, and apart from that to do what we want sufficiently determinate contents and weights for the purposes of coordination. The progress that we make would, however, teach us something extremely important, namely where the line is to be drawn between ethics as an a priori philosophical enterprise and an a posteriori empirical enterprise. The clear recognition that this line needs to be drawn and how to find it is the third payoff (for a

discussion of the impact of this line of thinking on disagreements between agents from different cultural traditions see Smith 2024a).

A fourth payoff of the argument from below is that it suggests the following plausible definitions of the standard deontic features. An action is morally wrong just in case there is a decisive moral normative reason not to do it; an action is morally permissible just in case it is not morally wrong; and an action is morally obligatory just in case there is a decisive moral normative reason to do it. What makes these definitions so plausible is that they explain why there is so much interest in the standard deontic features of action. Someone who consistently acts wrongly is someone who cannot be trusted not to do what they have a decisive categorical normative reason not to do. Given that other people helping but not interfering is part of what makes it possible for each of us to live the kind of life we want to live, which is just the kind of life that we have hypothetical reasons to live, it is therefore to be expected that we would have a special interest in identifying those who cannot be trusted to do their part in this.

A fifth payoff is connected to what I earlier called the guiding idea of *TMP*, the idea that folk morality is committed to the truth of moral rationalism. As I said, moral rationalism implies that moral facts aren't just, if known, known a priori, but are known to ideal reasoners. The argument from below delivers on that implication of moral rationalism, thereby ensuring that moral facts can play the role of making it possible for everyone to get along by freely doing what they have most normative reason to do. That is, after all, exactly what the community of ideal agents do. Moreover the argument from below shows that it is not absurd to hope that, in the fullness of time, non-ideal agents like ourselves will do the same thing.

References

- Bukoski, Michael 2016: "A Critique of Smith's Constitutivism" in *Ethics* (127) pp. 116–146.
- Dancy, Jonathan 1994–5: "Why There Is Really No Such Thing as the Theory of Motivation" in *Proceedings of the Aristotelian Society* (95) pp. 1–18.
- Gauthier, David 1987: *Morals By Agreement* (Oxford: Clarendon Press)
- Harman, Gilbert 1975: "Moral Relativism Defended" *Philosophical Review* (84) pp. 3–22.
- Haslanger, Sally 2012: *Resisting Reality: Social Construction and Social Critique* (New York: Oxford University Press).
- Horgan, Terence and Mark Timmons 1996: "Troubles for Michael Smith's Metaethical Rationalism" *Philosophical Papers* (25) pp. 203–231.
- Howard, Nathan and Mark Schroeder 2024: *The Fundamentals of Reasons* (Oxford: Oxford University Press)

- Joyce, Richard 2001: *The Myth of Morality* (Cambridge: Cambridge University Press).
- Kelly, Thomas and Sarah McGrath 2010: "Is Reflective Equilibrium Enough?" *Philosophical Perspectives* (24) pp. 325–359
- Lenman, James 1999: (1999) "Michael Smith and the Daleks: Reason, Morality, and Contingency" *Utilitas* (11). pp. 164–177
- Lewis, David 1984: "Putnam's Paradox" in *Australasian Journal of Philosophy* (3) pp. 221–236
- McGrath, Sarah 2010/2013: "Moral Realism without Convergence" *Philosophical Topics* (38) pp. 59–90.
- Nagel, Thomas 1970: *The Possibility of Altruism* (Princeton: Princeton University Press)
- Parfit, Derek 2011: *On What Matters* Volume One (Oxford: Oxford University Press)
- Pettit, Philip and Michael Smith 1990: "Backgrounding Desire" in *The Philosophical Review* (99) pp. 565–592
- Portmore, Douglas 2011: *Commonsense Consequentialism: Wherein Morality Meets Rationality*. (Oxford: Oxford University Press).
- Scanlon, Thomas M. 1998: *What We Owe to Each Other* (Cambridge: Harvard University Press)
- Smith, Michael 1994: *The Moral Problem* (Oxford: Wiley–Blackwell).
- Smith, Michael 2009: "Two Kinds of Consequentialism" in *Philosophical Issues* (19) Metaethics, pp. 257–272
- Smith, Michael 2012: "A Puzzle about Internal Reasons" in *Luck, Value and Commitment: Themes from the Ethics of Bernard Williams* edited by Ulrike Heuer and Gerald Lang (Oxford: Oxford University Press, 2012) pp. 195–218.
- Smith, Michael 2020: "The Modal Conception of Ideal Rational Agents: Objectively Ideal Not Merely Subjectively Ideal, Advisors not Exemplars, Agentially Concerned Not Agentially Indifferent, Social Not Solitary, Self-and-Other Regarding Not Wholly Self-Regarding" in *Explorations in Ethics* edited by David Kaspar (New York: Palgrave Macmillan), pp. 59–79
- Smith, Michael 2021: "The State of Nature" in *Oxford Studies in Normative Ethics* edited by Mark Timmons (Oxford: Oxford University Press) pp. 270–289
- Smith, Michael 2024a: "Clashes of Culture" in *Human Minds and Cultures* edited by Sanjit Chakraborty (Basel: Springer Nature Switzerland) pp. 73–88.

- Smith, Michael 2024b: “The Revised Moral Problem” in *Filosofiska Notiser*, (10) 2023 pp. 207– 216.
- Stocker, Michael 1979: “Desiring the Bad: An Essay in Moral Psychology” *Journal of Philosophy* (76) pp. 738–53
- Thomson, Judith Jarvis 2010: *Normativity* (Chicago: Open Court Publishing Company).
- Williams, Bernard 1981: “Internal and External Reasons” in his *Moral Luck* (Cambridge: Cambridge University Press).