BELGRADE PHILOSOPHICAL ANNUAL/
FILOZOFSKI GODIŠNJAK 37(1)/2024
Institute for Philosophy, Faculty of Philosophy,
University of Belgrade
Belgrade, Čika Ljubina 18–20

Belgrade

Year XXXVII
YU ISSN 0353-3891
UDK-1

Belgrade Philosophical Annual is published twice a year and is available online at http: //www.f.bg.ac.rs/bpa/index.html

Printed by
Službeni glasnik, Belgrade

This issue is financially supported by
Ministry of Education, Science and Technological Development of the Republic of Serbia

The statement on publication ethics can be found at the journal website http://www.f.bg.ac.rs/bpa

BELGRADE PHILOSOPHICAL ANNUAL 37(1)/2024

THE MORAL PROBLEM:
REFLECTING ON 30 YEARS OF PHILOSOPHICAL INQUIRY

THE MORAL PROBLEM: REFLECTING ON 30 YEARS OF PHILOSOPHICAL INQUIRY

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

 DOI: 10.5937/BPA2437007S

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 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 flukey. 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).

Frank Jackson
Australian National University
frank.jackson@anu.edu.au

Original Scientific Paper
UDC 17.022.1
111.6
Received: July 30, 2024
Revised: November 01, 2024

Accepted: November 11, 2024

THE MORAL PROBLEM: A CORRECTION TO THE KEY THOUGHT

Abstract

I argue that the three drugs example makes trouble for the role Smith gives to being fully rational in his solution to the moral problem, given his understanding of what it takes to be fully rational. I conclude by suggesting he might have drawn on a different understanding of what it takes to be fully rational.

Key words Fully rational · three drugs example · Bernard Williams

1. Preamble on the key thought

There are: the actions we desire to perform; the actions we desire to desire to perform; the actions idealised versions of ourselves would desire to perform after reflection; the actions we humans have evolved to desire to perform once we became members of mutually supportive communities; the actions it would be rational to desire to perform; and so on and so forth. This observation prompts the following thought: can we give an account of what it takes for an action to be what one morally ought to do, to be the right thing for one to do, in terms of the action's having some desirability property or other? This thought could be fleshed out in a variety of ways, depending on how the relevant desirability property is cashed out. I listed some of the possible ways a sentence or two ago.

Michael Smith in *The Moral Problem* offers us a version of this important thought, where 'what it takes for an action to be the right thing to do' is to be understood as a kind of analysis of being the right thing to do. This essay is about a problem I see for his version of the thought, and about how to modify it to avoid the problem. Once modified, should we endorse Smith's account of what it takes for an action to be what one ought to do? As it happens, I favour a view according to which what it takes for an action to be what one ought to do turns *in part* on its connection to desire, but on much else besides.[1] However, that's a question for another time.

1 For a recent version of this view, see Jackson and Pettit (2023).

 DOI: 10.5937/BPA2437033J

The problem for Smith's version of the thought arises from the conjunction of two features of it.[2] I will start by detailing the two features in question.

2. Rational desires and motivation

I start with the rational desire feature. In Smith's terms, what one ought to do is what is *desirable*, where what is desirable is what it is rational to desire, or, as he sometimes puts it, rightness is the feature we would want actions to have if we were *fully rational*. Here the notion of being fully rational is playing a key role and Smith has, as you would expect, a number of comments designed to explicate the notion. For my purposes, we can finesse much of his discussion. What is important for the argument to come is that a necessary condition for a feature to be one we would want an action to have if we were fully rational is that the feature is one we would want an action to have if we had 'all relevant true beliefs (156)'.[3] Once upon a time it was rational to desire margarine on one's bread instead of butter in the following sense: the evidence available back then was that margarine was much healthier than butter. We now know that this is wrong. It's much of a muchness. (It is, I understand, olive oil that's much better.) Was it fully rational *back then* to desire margarine ahead of butter? Not in the sense Smith has in mind. Had we had all true relevant beliefs back then, we would not have desired margarine ahead of butter. A necessary condition for a desire to be fully rational in Smith's sense is that it be what one would desire if one had all the relevant facts to hand. We might call this the full (relevant) information requirement.

Now for the motivation part of his account. Smith has a particular reason for liking an account of rightness in terms of what it would be rational to desire. He sees it as a key plank in explaining how believing that an action is the right thing to do can be appropriately connected to motivation. The exact way to articulate the connection on Smith's account is a contested matter. I expect that some contributions to this symposium on Smith's book will focus on this issue and, more generally, on the question of how the belief that an action is the right thing to do is connected to being motivated to perform that action, and indeed on whether this belief is a belief strictly speaking. We can, however, set these matters to one side. What is important for what is to come is the basic, and surely interesting and attractive idea, that Smith finds so appealing. Here is one way of expressing matters – in my phrasings, not his. If someone believes that an action is such that if they were fully rational, they would want to perform that action, then what they believe points towards, in some good sense, performing the action in question. We can see, at least in the broad, how a belief about the nature of an action, albeit a special kind of belief about its nature, might, in and of itself,

2 I should highlight that in what follows I am discussing the view as it appears in *The Moral Problem*; I do not address subsequent developments.

3 He is quoting from Bernard Williams's conditions for being fully rational (Williams 1980), but Smith makes it clear that it is a condition he accepts and gives his reasons in the pages that follow 156.

recommend performing the action. After all, if one refrained from performing the action in question, one would be failing to do that which one believed was the fully rational thing for one to want to do. One might well be tempted to say that even if Smith hasn't given certain theorists exactly what they want from the connection between believing that an action is right and being motivated to perform the action, he has given them something near enough and, maybe, all there is to give. Not every pre-analytic intuition can be saved.

In sum, Smith is offering us an account of what it takes for an action to be right that allows us to make sense of the motivational role of the belief that an action is right, for, according to him, in believing that X is what one ought to do, one believes that X is that which it is fully rational for one to want to do. That's his key thought. Of course, there are all sorts of qualifications and refinements, but they do not matter for what follows.

Now for the problem. I will introduce it via a discussion of a well-known example.

3. The three drugs example[4]

Suppose that Jane Doe is treating John Doe for a serious but not life-threatening disease. She has to decide between administering drug A, drug B or drug C. They are her options in the sense that she knows that these are all the relevant actions that are within her power. She is all but certain that one or other of drug A or drug C would effect a complete cure but that the other would kill John. She justifiably assigns to each a 50% chance of being the killer drug and a 50% chance of being the drug that would completely cure him. She also justifiably believes that it is very likely that giving drug B would effect a partial cure. She is rightly confident that there is no way, or no way in the time available, for her to obtain extra information bearing on her decision about which drug to administer. Finally, as a matter of fact, drug A is the one that would completely cure him.[5]

What ought Jane to do? I mean this question to be understood as a question about what she ought *morally* to do. In prudential terms, it is uncontroversial what she ought to do; she ought to administer drug B. She would likely be disciplined by the medical authorities if she didn't. But the answer to our question – the question of what she ought morally to do – is a controversial matter. This is a matter of record, for there are three live answers – live in the sense of being answers that have significant support among the many who have thought seriously about cases of this kind – to our question.

4 Modified from Jackson (1991), but many describe cases of this general kind.

5 It would be a mistake to think of cases like these as discoveries of philosophers. Financial advisors have long known that the shares they ought to tell their clients to buy are very often shares that are certain not to deliver the best returns. The shares that will deliver the best return are instead one or another of various highly speculative ones. The advisors' problem is that these highly speculative shares also include the shares that will deliver the worst returns, and there is no way for them to tell the sheep from the goats.

One answer is that Jane ought to administer drug B. This is the answer I favour. It is the answer typically favoured by those with an interest in decision theory. A second answer is that Jane ought to administer drug A. She doesn't and cannot know this of course, but, say supporters of this answer, why suppose that what one ought to do is always knowable? A third answer is to distinguish two senses of 'ought'. In the objective sense of 'ought', Jane ought to administer drug A; in the expective (often called the prospective) sense of 'ought', she ought to administer drug B. I trust the labels explain themselves.

The third answer, as we are understanding it here, goes beyond simply providing labels for administering drug A and for administering drug B – something one could hardly complain about. It insists that this is *all there is to say* about the question of what Jane ought to do. Its supporters hold that there is nothing to disagree about when addressing the question of what Jane ought morally to do. For we can all agree that in one sense, the expective sense, she ought to administer drug B, and in the other sense, the objective sense, she ought to administer drug A.

I think we should set this 'quietist' answer aside. As a supporter of the first answer, I insist it is *immoral* to take unjustified risks with people's lives even in cases where one gets way with it, as would be the case for Jane if she administered drug A. It is something one ought not to do! But if the supporters of the third answer are right, this is a nonsense position to take. There is nothing to quarrel about. This is very hard to believe. Supporters of the second answer are equally unhappy with the quietist answer. When they insist that the right thing for Jane to do in our example is settled by the facts of the case, not her rational credences, they are taking a position in ethics, not a position on how to use words.

Here is a way to capture what is wrong with the quietist position. There are famous cases – too famous to need detailing here – where the deontological answer concerning what one ought to do in them differs from the consequentialist answer. These cases, and discussions of them, are part and parcel of the ongoing debate in ethics between deontologists and consequentialists. We might coin two terms – 'ought-d' and 'ought-c': the first for the deontological answer concerning what one ought to do in these cases, the second for the consequentialist answer concerning what one ought to do in them. We could then say that, for each of the famous cases, there is what one ought-d to do and what one ought-c to do. Would the right response to this exercise in labelling be that we have shown that there is no substantive disagreement between deontologists and consequentialists concerning what ought to be done in these famous cases, on the basis that all parties agree about the ought-d answer and about the ought-c answer concerning what ought to be done in them? Surely not!

In what follows, I am going to presume that the two viable views about what Jane ought to do in the drugs example are: to administer drug B, or to

administer drug A. What we will see is trouble for Smith from both answers. For both, there is trouble arising from his requirement on what it would be fully rational to desire, the requirement that it be, amongst other things, what we would desire if we had all the relevant information.

4. The trouble if the first answer is correct

Suppose that Jane ought to administer drug B. This means, according to Smith, that her administering drug B is desirable, where this comes to its being what it is fully rational for her to want to do. But it isn't. To be fully rational is to be, amongst other things, what would be desired given all relevant information, and what would be desired by her given all relevant information is that she administers drug A, not drug B. If the first answer is correct, Smith's account of what it takes to be the right thing to do is mistaken. We will now see that if second answer is correct, there is trouble for his account of how a belief that an action is right might motivate one to perform the action.

5. The trouble if the second answer is correct

Suppose that Jane ought to administer drug A. She won't, however, do so in fact. We may suppose that she is a conscientious, principled member of the medical profession and, as such, we know that she will administer drug B. What is her motivation for doing so? We mentioned earlier that there is a prudential reason for administering drug B: she might well be subject to disciplinary action if she doesn't, and she will know this. But it would be an unwarranted slur on the medical profession to suppose that the only reason its members administer drug B in the kind of situation we are talking about is their fear of being disciplined if they don't. We know this because we know that most of them would administer drug B even if they happened to know that if they did not, they would escape censure. What then is Jane's (non-prudential) motivation for administering drug B?

The obvious answer, and in any case one possible answer, is that Jane believes that the morally right thing for her to do is to administer drug B, that she accepts the first answer to the question posed by three drugs example. Of course, if the second answer is correct, this belief of hers will be a false belief, but it will be a belief she has all the same, and it will be the belief that motivates her to administer drug B.

How, on Smith's account of what it is to believe that an action is the right thing to do, could this belief motivate her to administer drug B? His story about motivation is in terms of belief about what she would want to do if she were fully rational. But if she were fully rational, what she would want is that she administers drug A, *not* drug B. That follows from the full information condition on being fully rational when combined with the fact that drug A is

the one that would effect a complete cure. *Moreover*, Jane won't believe that if she were fully rational in the sense Smith has in mind, she would want to administer drug B. She will believe that she would want either to administer drug A or to administer drug C, for she knows that it is one of those two that would effect a complete cure, and she will know that the one drug she will not want to administer is drug B.

6. A way out

The trouble stems from his full information requirement on being desirable, on being that which it is fully rational to desire. It has always puzzled me why Smith included it. Surely it can be dropped without damaging his key insight.[6]

References

Jackson, F. (1991). Decision–Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics* 101 (3), pp. 461–482.

Jackson, F., & Pettit, P. (2023). Moral Functionalism, in *The Oxford Handbook of Moral Realism*, (ed.) Bloomfield, P., & Copp, D., Oxford University Press, pp. 246–263.

Williams, B. (1980). Internal and External Reasons, reprinted in *Moral Luck*, Cambridge University Press, 1981.

6 I have discussed the issues bruited here with many colleagues and friends, and owe special thanks to Michael Smith and Philip Pettit.

Joshua Gert
William & Mary
jngert@wm.edu

Original Scientific Paper
UDC 17.021.1/.2
17.034:165.6
17 Смит М.
Received: June 21, 2024
Revised: July 22, 2024
Accepted: July 25, 2024

MORAL REASONS AND *THE MORAL PROBLEM*[1]

Abstract
When Michael Smith published *The Moral Problem*, he advocated only Weak Moral Rationalism: the view that moral requirements always provide us with reasons that are relevant to the rationality of our action. But in the intervening years he has changed his position. He now holds Strong Moral Rationalism: the view that moral requirements are all-things-considered rational requirements. In this paper I argue that his change in view was motivated by two things. The first is his correct view that acting as one is morally required to act is never irrational. The second is what David Copp has called The Unitary View of Reasons: the idea that there are both moral reasons for action, and non-moral ones, and both sorts count as reasons that determine what is rational to do. This combination of views pushes Smith to hold that an act counts as morally required just in case the moral reasons that favor it outweigh all other reasons, both moral and non-moral. But, I argue, there is an attractive position between Weak and Strong Moral Rationalism, which I call Moral Permissibilism. On such a view, moral requirements, while not always rational requirements (as against Strong Moral Rationalism), are always rationally permissible (as against Moral Anti-Rationalism). In order to advocate this common-sensical position, however, one must abandon the Unitary View of Reasons, and recognize that reasons of different kinds contribute to different kinds of normative verdicts.

Keywords: Michael Smith · *The Moral Problem* · Moral Rationalism · Normativity · Reasons

1. Introduction

In the first ten years after *The Moral Problem* came out, two things happened. The first was that the book entered the canon as a classic of metaethics, setting the agenda for a huge secondary literature. The second was that it became fairly

1 This paper was completed while I was in residence at the Center for Advanced Studies at the Norwegian Academy of Science and Letters. I would like to thank Mathea Sagdahl and Attila Tanyi for inviting me to participate in their project there, and to Caj Strandberg for extremely helpful written comments on an earlier draft, as well as to an audience at the University of Oslo.

settled opinion that Michael Smith was a strong moral rationalist: a defender of the thesis that moral requirements are also rational requirements.[2] Neither of these things was especially surprising. The first was a natural consequence of the book's great clarity, ambition, and originality. The second could be laid at the feet of Smith's idiosyncratic use of the phrase 'categorical requirement of rationality' in asserting that 'our concept of a moral requirement thus turns out to be the concept of a categorical requirement of rationality'.[3] That certainly sounds like the sort of thing only a rationalist would assert! Still, in using this phrase Smith only meant to be stressing that moral requirements entail reasons that are *desire-independent*.[4] But those reasons were still only *pro tanto* entities. So if one took the trouble to get clear on the way in which Smith used the relevant terminology, it did not seem right to me to interpret him in the standard rationalist way. In 2008 I therefore made the case for a different reading.[5] Although I offered a number of arguments in favor of that reading, the most persuasive lines in the paper should probably have been the explicit endorsement, by Smith, of the following position:

> That in some conflicts between morality and self-interest, either option is rationally permissible, and that this happens not merely when one of the options is morally supererogatory, but in cases in which the choice is between a genuinely immoral and selfish action, and a morally required one.[6]

Somewhat disappointingly – given the explicit endorsement from Smith – there was virtually no uptake of this paper, and the misreading of Smith

2 See Copp (1997), Horgan and Timmons (1996), Lenman (1999), Sadler (2003), Swanton (1996), and Noordhof (1999). Strong moral rationalism can be contrasted with weak moral rationalism. According to the latter, moral requirements always provide reasons for action, though one may sometimes be rationally permitted to act against those reasons. I will always use 'moral rationalism' as a label for the stronger position.

3 Smith (1994), p. 87.

4 See, for example, Smith (1994), p. 185. The same mode of expression persists even now, though I am less sure about how to interpret it. For example, in one of a series of lectures given at Nanjing Normal University and Shandong University in November of 2018, Smith says 'moral requirements are themselves "categorical" requirements of reason, *which is to say* that they are reasons to act that agents have simply in virtue of their being rational' (my emphasis). But, again, reasons are naturally understood as *pro tanto* entities that can be overridden by other reasons.

5 Gert (2008).

6 Gert (2008), p. 1 n. 1. See also Smith (2002), in which he makes the following claim:

> [Bernard] Gert is right to emphasize that there is a category of rationally allowed acts alongside those that are rationally required, and I think that he is also right to insist that many so-called conflicts between morality and self-interest are best thought of in this sort of way. Agents have a free choice to decide in which way they will act, at least within certain limits. They are not rationally required to act in the one way or in the other. Moreover, I think he is right that a theory of rational action that suggested otherwise would be flawed in a quite decisive way.

persisted. Time, however, seems to have turned that mistaken view of Smith's position into a correct one. That is, Smith – in the non-dogmatic spirit found in the very best philosophers – has reconsidered his position.[7] He now explicitly endorses the rationalism he formerly denied. As he now puts it, his view includes the following:

> an act's *being morally forbidden* entails that there are moral reasons, that some of these are reasons not to perform the action, and that these moral reasons are weightier than the moral or non-moral reasons to perform the action.[8]

and

> each of us would choose to act morally if we had and exercised the capacity to respond rationally to the circumstances in which we find ourselves.[9]

In this paper I will not repeat the arguments I offered earlier, though they still seem to me to be correct as an interpretation of *The Moral Problem* itself. Smith simply has a different view now. Rather, it is the point of the present paper to try to understand what motivated his change in view, and to assess whether those motivations were good ones. Other contributors to this volume will, I think, address themselves to Smith's transition to an explicit constitutivism about reasons. My own focus has to do with the way in which Smith has changed his view of the way in which moral reasons determine moral status.

2. The *Unitary View* of Reasons

One position that Smith has held consistently from 1994 to the present is that there are both moral reasons for action, and non-moral ones, and that both sorts count in determining what we have all-things-considered reason to do. More generally, he endorses what, following David Copp, we can call the *unitary* view:

> all (genuine) reasons are reasons *simpliciter* or *unqualifiedly*. Different kinds of consideration 'give rise' to different kinds of reason, but the unitary view insists that the status of a consideration as a reason does not depend on its being considered in relation to any particular normative system.[10]

The idea is that 'what we have normative reason to do all things considered is a matter of the relative strengths of the pro tanto normative reasons we have',

7 I am thinking of Wittgenstein, Putnam, and Foot, among others.

8 Smith (2018), p. 48.

9 Smith (2013), p. 9.

10 Copp (2009), p. 24.

where those reasons might be moral or non-moral.[11] Smith also holds that a reason is 'a consideration that can rationally justify our acting in a certain way'.[12] So we can also say that, on his version of the unitary view, what one is rationally justified in doing is determined by the weights of *all* the normative reasons – of whatever substantive kind – that bear on one's action.

In *The Moral Problem* the distinction between moral and non-moral reasons was drawn in terms of what Smith called 'platitudes' about their content, though no particular way of drawing the distinction was defended.[13] A couple of suggestions – only meant to get us in the ballpark – were that moral reasons have to do with promoting human flourishing and expressing equal respect. Smith also held that in some cases moral reasons outweigh non-moral ones, while in other cases the reverse is true, and in still other cases – which we might call cases of rational underdetermination – neither outweighs the other.[14]

The unitary view has one significant theoretical cost, no matter how it is developed: it makes hard to see why we should have any distinctive concern at all with the particularly *moral* status of an action. After all, what really matters, on this view, is whether an action is maximally favored by the totality of reasons. Practically speaking, it seems unimportant whether those reasons are moral, or non-moral, or a mix. Consider a particular action, φ, and suppose the following:

(i) φ is uniquely maximally favored by the totality of reasons, so that φ is uniquely rationally justified.

(ii) The moral reasons that favor φ outweigh all the opposed reasons, both moral and non-moral, combined.

On one view of moral requirements, (ii) entails that we can say that φ is morally required. But (ii) seems to be a matter of purely academic interest – using 'academic' in the pejorative sense. The fact that (ii) is true seems no more relevant to any practical matter than would be either of the following:

(iii) The short-term-consequence-related reasons that favor φ outweigh all the opposed reasons, both those that have to do with short-term consequences, and all other reasons, combined.

---

11 Smith (1996), p. 167.

12 Smith (1996), p. 164.

13 Smith (1994), p. 183-4.

14 For example, Smith (1994, p. 183) quotes Wolf (1982) approvingly, from a paper in which she argues against moral rationalism (among other things). Admittedly, the claims Smith makes in connection with Wolf only strictly entail that moral reasons can conflict with other important reasons – not that they can ever lose when they conflict in this way. Still, twenty years ago Smith pointed out his approval of Wolf to me in order to explain his great surprise that anyone would ever have taken him to be a strong moral rationalist.

(iv) The child-related reasons that favor φ outweigh all the opposed reasons, both those that have to do with children, and all other reasons, combined.

Neither (ii), (iii), nor (iv) make any difference to whether we ought, all things considered, to perform φ or not: the truth of (i), all by itself, already settles that.

Given the above, we should expect that someone with the unitary view of reasons will not much care what criteria are used to classify reasons as *moral*; after all, it doesn't really matter. And, indeed, at the time he wrote *The Moral Problem*, and for some years after, Smith was not concerned to specify the criteria, except to get us 'in the ballpark'. For example, when it came to determining whether a desire for the welfare of one's own family corresponds to a *moral* reason to promote their welfare, his view was that there was no determinate answer. Moreover, he did not think we gained anything by legislating one way or the other, precisely because – as I've been emphasizing – what really matters is the relative strengths of one's reasons. It is a matter of indifference which subset of these reasons is classified as moral.[15]

Suppose that despite its lack of practical importance, an advocate of the unitary view wanted to explain what is distinctive about moral reasons. One very popular strategy for doing so will not be available. In particular, the unitary view is in tension with any attempt to explain what moral reasons are by reference to a moral theory that appeals to practical rationality as an explanatorily more basic notion. Suppose, for example, one held morality to be the system that it would be *practically rational* to put forward under certain circumstances (say, behind something like the veil of ignorance). There would then be something quite fishy about explaining moral reasons in terms of the rules of that moral system. After all, on the unitary view what counts as practically rational depends to some degree on what moral reasons there are. So we would have to know what moral reasons there are before we could construct the moral theory that explained what moral reasons there are.

How, then, will an advocate of the unitary view explain what moral reasons are? One strategy would appeal to their formal features – such as universalizability. But, as contemporary critics of Kant continue to note, an appeal to purely formal features seems unlikely to yield any determinate substantive content, let alone a content that we would recognize as distinctively moral.[16] Given worries about characterizing moral reasons in a purely formal

15 See Smith (1999), pp. 56–57. Smith (2018) now holds that for a reason to be moral it must be impartial (characterizable as features of agents as such), and must not depend on any contingent desires of the agent. This seems to rule out *de se* reasons of the sort that he (1994, pp. 169–71) was formerly happy to include among the full set of reasons. The 2018 and 1994 positions both contrast with one he held during a middle period, according to which all moral reasons were agent-relative (see Smith 2011, esp. pp. 360–1).

16 Smith himself (2010, p. 130) seems to agree that Kant's universalization argument fails, on its own, to establish substantive normative requirements.

way, a direct appeal to the *substantive content* of moral reasons may seem more promising. But it is worth noting that the distinction in content between self-interested and altruistic reasons is not up to the job. As has often been noted, a great deal of the most egregiously immoral action is favored by reasons that cannot plausibly be construed as self-interested. Indeed they often involve significant personal risks, and are rationally justified because of benefits they provide to other people: to family and friends, for example. Those reasons do not provide any moral justification, despite being altruistic.

In more recent work, Smith has defended a view of moral reasons in terms of a substantive content that stems from the constitutive nature of agents.[17] Agents are, by their very nature, beings with the capacity to *act*: to form beliefs about the world, and to realize their desires in light of those beliefs. And they can exercise this capacity in better and worse ways. I won't rehearse Smith's argument here, but the upshot is that ideal agents will all have certain *overriding* desires. Among these will be desires to avoid interfering with or undermining anyone's agency. And he thinks *these* desires correspond to our moral reasons. This is not the place to present or criticize the details of this newer more Kantian view.[18] Rather, my focus is on the unitary view of reasons that he continues to endorse, and on an alternative that I think is preferable. But before getting to that more preferable alternative, let me discuss two ways of incorporating the unitary view into a view of moral obligation and permission.

2.1 The Compartmentalizing *View of Moral Obligation and Permission*

If one has the unitary view of reasons, then one tempting view of moral obligation is that it is determined entirely by the relevant moral reasons: non-moral reasons are beside the point. We can call this the *compartmentalizing* view. According to the compartmentalizing view, for an act to count as *morally required*, it must be the unique act that is maximally favored by the balance of specifically moral reasons. To put it compactly, we can say that such an action is *uniquely maximally morally favored*.[19] If there is a set containing a number of actions available to the agent, none of which is *less* favored by the balance of moral reasons than any other available action, then no action will

17 Smith (2013).

18 See Enoch (2020) for a critical discussion that also presents some of Smith's potential counterarguments.

19 More precisely this should be put in terms of acts-under-descriptions. A *specific* act of returning a book – in a certain way, at a certain time, wearing certain clothes – might not be morally required. Still, one might be required to return the book. What this means is that only acts that fit the 'returning the book' description can be found in the set of maximally morally favored actions. In returning it in the specific way one did, one therefore counts as doing what was morally required, though one could have done so in a different way.

be *uniquely* maximally morally favored.[20] Still, all the actions in that set will be *maximally morally favored* – just not uniquely so. Any action in that set will count as *morally permissible*. For an action to count as *morally wrong*, it must not be morally permissible. That is, there must be some moral reasons relevant to it (so that it has a moral status at all) but some other act must be more strongly favored by moral reasons.

The compartmentalizing view entails that, technically speaking, actions to which no moral reasons are relevant (perhaps the options open to Robinson Crusoe, prior to his meeting Friday) will be neither morally permissible nor morally wrong. Still, there are plenty of non-moral reasons, and they determine an overall normative status. To understand how this determination works, we can use a classification scheme just like the compartmentalizing view, but setting aside the qualifier 'morally'. Such a view will take *all* the reasons into account – both moral and non-moral – and these reasons will determine whether an action is uniquely maximally favored, *simpliciter*, or in a set of actions that are all maximally favored, *simpliciter*, or not maximally favored, *simpliciter*. We can label actions of the first sort 'rationally required', actions of the second sort 'rationally permissible', and actions of the third sort 'irrational'.

The compartmentalizing picture is certainly a coherent one. But it has implications that are controversial. One mildly controversial implication is that some rationally permissible actions might also be immoral. That is, an action that is *not* maximally favored by moral reasons might nevertheless be maximally favored by the totality of practical reasons, since that totality includes non-moral reasons as well. This is a controversial consequence: it allows – as Smith used to allow – that immoral behavior can sometimes be rationally permissible. I myself think that, though controversial, this is actually correct. But there is a worse consequence not far off: an action to which moral reasons are relevant, and which counts as immoral because it is not maximally favored by moral reasons, might nevertheless be *uniquely* maximally favored by the totality of practical reasons once we take non-moral reasons into account. In that case, the uniquely rational thing to do would be something that is immoral. We would be rationally required, in such a case, to act immorally. Given the role that practical rationality tends to play in moral and political philosophy, this would be a disaster. Putting it in the terms Smith employed in *The Moral Problem*, it would mean that the person best placed to give us advice – our fully rational selves – would advise us not to act in the morally required way. This is worse than the denial of moral rationalism, when that view is understood as the view that moral behavior is rationally required. It is the view that moral behavior is sometimes rationally prohibited.[21] Later I will call this view moral anti-rationalism.

---

20 I use 'not less favored' instead of 'equally favored or more favored' to make room for incommensurability or parity. Smith seems open to these theoretical possibilities.

21 It is clear that Smith, at a certain point, also regarded this as a disastrous theoretical consequence. See Smith (2013), p. 12, where he expresses the worry that Humean theories might seem to have it as an entailment.

2.2 The Preponderance *View of Moral Obligation and Permission*

One response to the problems I've just described with the compartmentalization view keeps the unitary view of reasons, but includes a distinct account of the relation between moral reasons and moral status. On this alternate account, for an act to be immoral is for the moral reasons that count against it to outweigh *all* the reasons that favor it, whether moral or non-moral. We can call this the *preponderance* view. With such an account of immoral action, it is easy to define morally permissible action: it is just the complementary class. And we can define an act as morally required if it would be immoral to omit it – or, equivalently, if it is the *only* morally permissible act.[22]

Like the compartmentalizing view, the preponderance view of moral status is certainly coherent. But it also yields controversial consequences. Assume, as Smith does, that we are rationally required to act on our weightiest reasons.[23] In that case, the preponderance view yields a strong form of moral rationalism, simply by definition: any immoral action will have moral reasons against it that are weightier than all the moral and non-moral reasons that favor it, combined. By itself this does not count decisively against it: moral rationalism is part of the legacy of Kant, and, though controversial, it has many defenders. But there is another consequence of the preponderance view that is more problematic: it is hard to see how it will end up classifying actions in a way that lines up with our pre-theoretical views of what is morally permissible, and what is not. Let me explain.

Recall, on the unitary view of reasons, there is a large and heterogeneous set of considerations that count as reasons for and against action. One such reason might be 'the act will cause me (the agent) pain'. Another might be 'the act will fulfill a promise to someone else'. Still another might be 'the act will save the life of a stranger'. Let us accept that some of these count as moral reasons, and others as non-moral. But they are all reasons, which means – given the unitary view – that they are all relevant to the rational status of any act to which they apply.[24] The problem I want to highlight stems from the fact that some of the non-moral reasons of relevance to an act seem to be *morally* relevant in a certain sense, while others do not seem to be. That sounds odd, so let me offer an example to make it clearer.

---

22 See Smith (2018), p. 48. Liz Harman – who endorses the unitary view of reasons – has a view of moral obligation that is distinct but related to the preponderance view. According to Harman (2020) we say someone morally ought to perform an action if the most salient reasons that determine that we ought to perform it turn out to be moral ones. It follows trivially that we ought to do anything that we morally ought to do. Still, Harman's view is not a version of moral rationalism, since moral reasons do not in themselves have any special status, or add up in any special way to determine what we ought, overall, to do.

23 Smith (1996, pp. 167–8).

24 For the sake of simplicity I am leaving aside considerations of probability. This omission does not affect the broader point.

It seems fairly uncontroversial that the fact that an act will spare me a great harm is *relevant to the moral status* of an action. After all, such reasons are often taken to provide me with a *moral* justification for doing what would otherwise have been morally prohibited. I am justified in breaking a promise to meet someone for lunch, for example, if I begin to have symptoms of a serious medical problem that would benefit from immediate attention, even if I would probably survive if I waited until after lunch before going to the hospital. The preponderance view captures such cases nicely, since when such self-interested reasons are sufficiently powerful, it is easy to understand how they might justify acting against the relevant moral reasons. The problem is that *some* non-moral reasons do not seem to be relevant to the *moral* status of an act in this way. For example, it just does not matter how much money I will pay you to kill my boss; killing her remains morally prohibited. We know this without having to determine the power of the reason provided by the money. It is simply *irrelevant* to the moral status – incapable of changing it – no matter how powerful it is. And the same is true even if I am paying you to merely to break my boss's legs, or to step on her foot, or to break the taillights on her car. If moral and non-moral reasons compete on the same field – as the preponderance view holds – the financial reason should provide moral justification in precisely the same way as the reason having to do with avoiding harm for oneself. But in many cases our strong pre-theoretical sense is that they are not even relevant. And yet the same financial reason could justify me in acting against certain reasons – even quite strong ones. For example, it might well be rationally permissible to suffer a great deal of pain for enough money. An advocate of the preponderance view needs to explain this. Of course, they might be able to, though the task does seem challenging. But the more general problems with the unitary view will persist. These were:

(i) Difficulties explaining why we should care about moral reasons and moral status at all, given that rational status is, in an important sense, the only status that matters practically.

(ii) Significant restrictions on the kinds of theorizing available, by which to specify the category of moral reasons.

3. The Relational View of Reasons

In contrast with the unitary view, it is possible to hold that moral reasons and non-moral reasons simply do not compete on the same field, because they do not belong to the same normative domain. That does not mean that we should deny that there is anything interesting to say about the relation between moral reasons and generic practical reasons; I will make a suggestion about this later on, and I will also allow that one and the same fact – say, that my act will prevent someone from being hurt – can be both a moral and a generic practical reason. But it does mean that we should reconceive what

it means to classify reasons as moral or generically practical. My suggestion is the following: a moral reason is a consideration that makes a *systematic contribution* to the moral status of an action, while a generic practical reason is a consideration that makes a *systematic contribution* to the rational status of an action. In contrast to the unitary view, and again following David Copp, we can call this the *relational* view of reasons.[25]

It is easy to see how one might extend the relational view of moral and generic practical reasons to explain the nature of legal reasons, aesthetic reasons, reasons of etiquette, at least to the extent that overall status in these domains depends, in a systematic way, on the contributions – pro and con – of the various relevant facts.[26] A worry that may arise at this point is that if all of these normative domains are distinct, we will be at a loss to answer the question 'What should I do, in these circumstances?'[27] After all, different domains may yield conflicting answers. An act might be legally required, but morally or rationally prohibited. And one might worry about cases in which a morally prohibited action is nevertheless rationally required, or in which a morally required action is rationally prohibited. How should we act when faced with conflicts like these?

It is well beyond the scope of this paper to provide anything like a full answer to the general question of how one should decide to act, given the possibility of conflicts in the verdicts delivered by distinct normative domains. But elsewhere I have defended the thesis that practical rationality is the *fundamental* normative notion applying to action.[28] By this I mean – in part – that if it is determined that an action is irrational, the question as to whether one should perform it is closed: one simply should not. Still, it is consistent with the fundamental normative role of rationality that in many cases one has a number of rationally permissible options. When that happens, one must exercise one's capacity, as an agent, to exercise one's will, and *choose*.[29] It is also consistent with the fundamental normative role of

25 Copp (2009, p. 24) puts it in the following way: '[a]ccording to [...] "the relational view", something is a reason only in relation to a given normative system'. In Gert (2007b) I described moral reasons, as characterized by the unitary view, as *moral generic reasons*, while I described moral reasons, as characterized by the relational view, as *systematic moral reasons*. And I noted that Smith understands moral reasons as moral generic reasons. I also implicitly understood him – I think correctly – to have been working with the compartmentalizing view, rather than the preponderance view. Recently he has moved to the preponderance view. This is what pushes him into a strong moral rationalism.

26 It is worth noting that not all normative domains need be of this sort. Personally, I do not think that overall aesthetic status is fruitfully thought of as determined by aesthetic considerations in any *systematic* way.

27 See Sagdahl (2022) for an extended discussion of the sort of normative pluralism I am describing, and also for discussion of this problem of choice.

28 Gert (2004).

29 I therefore endorse what Joseph Raz (1997) describes, favorably, as 'the classical conception' of human agency.

rationality that rationality never requires immoral behavior. In particular, it may turn out that there is *always* sufficient generic practical reason to perform any morally required action. In fact, if we take 'sufficient' here to mean 'sufficient to *justify*', then I think that there is an argument that this is the case. I will sketch that argument in the final section of the paper.

On the relational view of reasons, one plausible moral reason might be 'the act is necessary to avoid causing someone a certain amount of a certain sort of pain'. As a moral reason, this consideration counts, morally, in favor of the act to which it is relevant. It is worth noting, though, that the *very same* consideration – that the act is necessary to avoid causing someone that particular kind and quantity of pain – might *also* count as generic practical reason: the sort of reason that determines overall rational status. And, in some legal systems, it might also count as a legal reason. Again, on the relational view, to determine if a consideration is a reason belonging to certain domain, one needs to determine whether that consideration makes a *systematic contribution* to whatever statuses are connected with that domain. Such statuses include, for example, the status of being legal or illegal, being polite or impolite, being rational or irrational, or being morally permissible, required, or supererogatory.

The contributions considerations make count as *systematic* if the overall normative status of an action within the relevant domain can be understood as the result of a *general function* of those contributions.[30] The simplest way for a consideration to make a systematic contribution would be for it to be associated with some kind of constant weight value, whether positive or negative. Simple desire-satisfaction utilitarians who think we can assign desires something like cardinal strength values are likely to endorse such a simple view of moral reasons. Moral consequentialists who are also pluralists about value are likely to make the picture more complex by countenancing relations of incommensurability or parity between instances of distinct values.[31] It is worth noting that the very same fact about an action might make *different* systematic contributions to the rational status of an act, as against its moral status, much as the very same fact about a drink – its alcohol content – might contribute differently to its capacity to intoxicate, as against its capacity to hydrate. For example, the fact that an act will hurt someone else might have a much greater impact on its moral status than on its rational status.

---

30 See Gert (2007a) and Berker (2007) for different arguments that we should understand reasons in terms of systematic contributions in this way.

31 Some moral particularists will deny that any substantive considerations make systematic contributions of this sort to moral status. As a result, there will be – if the view of reasons offered here is correct – no moral reasons. And some more extreme particularists will make the same claim about rationality and – therefore – generic practical reasons. The former view deserves some consideration: moral status may not be a function of reasons. I do not think that particularism about practical rationality, however, has much plausibility.

I myself have argued that, at least when it comes to the domain of practical rationality, the contributions that reasons make can vary along two dimensions. A generic practical reason can count (i) towards the status of an action as required, and (ii) towards its status as justified.[32] That is why it makes sense to ask, of any such reason, both 'How much can this reason rationally require me to sacrifice?' and 'How much can this reason justify me in sacrificing?' These questions are distinct. Justification is a matter of contributing to the *permissibility* of an act. Some considerations, in some domains, can provide a great deal of justification – can make it permissible to do things that would otherwise be very strongly prohibited – without having the power to require anything. For example, considerations of self-preservation seem to act this way in the moral domain. Consider: the fact that an act is necessary to save my life can *morally justify* very many acts that would otherwise be significantly immoral, even if it cannot justify *every* such act. So we can say that it has very significant moral justifying strength. But, arguably, I am not morally required to act in self-preserving ways.[33] If so, we can say that this moral reason has no moral requiring strength.

My own view is that some considerations that are moral reasons – for example, that an action is necessary to avoid causing a stranger some significant pain – have both requiring strength and justifying strength in the moral domain while only having justifying strength in the domain of practical rationality. As I will discuss in the final section of the paper, this helps explain why one is *always rationally justified* in acting in morally required ways, even though one is *not always rationally required* to do so. This is a view that Smith used to hold. But he can no longer hold it, since – under pressure to avoid saying that some morally required action is irrational – he has moved from the compartmentalization view to the preponderance view of moral requirements. And the preponderance view entails a strong moral rationalism.

It should be stressed that it is only *basic* or *canonical* reasons that must make systematic contributions to normative status. Other considerations can count as *derivative* reasons. For example, the fact that the pastry you gave me is made with almond flour is a reason for me not to eat it – given that I am allergic to nuts. But it is only a derivative reason. It counts as a reason in this derivative way because it explains why eating the pastry would cause me a great deal of discomfort. It is – at least plausibly – this latter fact that is a basic generic practical reason. It *always* counts against the rationality of action in the same way. If the normative status of an action is a function of the relevant reasons, it is only a function of basic or canonical reasons. Adding derivative reasons would result in over-counting: we would have to count the fact that

32 See Gert (2016) for an overview of a number of distinct arguments for this conclusion.

33 More precisely, when I am morally required to do so, it is because others are depending on me, or I have made a relevant promise; it is not the result of a basic moral requirement of self-preservation.

the pastry contains almond flour, and *also* the fact that eating it would cause me so much discomfort.

Smith, in endorsing the preponderance view – which includes the unitary view of reasons – has an easy explanation of the content of morality in terms of rationality. And this will be true no matter how he decides to categorize a reason as a specifically moral one. Moral requirements will simply be a certain subset of rational requirements: the ones in which the moral reasons that favor the action outweigh all opposed reasons, both moral and non-moral. But the relational view of reasons can also explain morality in terms of rationality, despite their being distinct normative domains. For example, the relational view makes room for a moral theory according to which the moral status of an act is determined by the system of rules that would be put forward by *rational people* under certain conditions. Such a view depends on the distinctness of morality and rationality, since it requires that we be able to characterize rational people and rational choices prior to determining the rules of a moral system and – therefore – prior to determining what might count as a distinctly *moral* reason. Similarly, a maximizing utilitarian about rationality might be able to put forward an account of morality as the system that it would be rational to try to educate people to adopt. Morality would then be quite different from rationality, on the plausible assumption that things would go better if people adopted a morality of rules than if they tried to maximize utility with each individual action they perform. The preponderance view, on the other hand, understands rationality partly in terms of moral reasons, so that it is incompatible with contractualist moral theories and system-utilitarian moral theories of the sort just described.

The relational view of reasons also makes room for informative explanations of the *distinctive* significance of moral assessments, as against rational assessments. I will not advocate any particular explanation here: most of the explanation will depend on what the correct moral theory turns out to be. If it is something like Mill's view of the moral, according to which for an act to count as morally wrong is for it to be the case that it would maximize happiness to punish it, we can immediately see the significance of a claim that an action is immoral. That is, the immorality of an act entails that it ought to be punished. If, on the other hand, some kind of contractualist view is correct, then something else of significance will follow: that we can say, truly, to anyone contemplating an immoral act, that *they themselves* would have put forward a rule prohibiting that act, if they were fully informed and rational. Pointing this out *could* have some persuasive force. Other moral theories will have other implications.

On the preponderance view moral requirements are a subset of rational requirements, and we have no clear way to say what the practical difference is between something being morally required and its being rationally – but not morally – required. Indeed, given the way that Smith now moves from a rational requirement not to interfere with one's own rational capacities to a

wider requirement not to interfere with *anyone's* rational capacities – basically by claiming that this extension is simply a case of treating like cases alike – it is hard to see why we should react differently to someone who harms themselves in this way, as against someone who harms someone else.[34] But this is far from our natural attitude. We try to get self-harmers psychological help, and do not feel anger towards them. We treat those who harm others quite differently.

Let us continue to suppose – just to have something concrete in mind – that contractualism is correct. Given this supposition, the relational view of reasons will determine what counts as a moral reason by inspecting the rules of the system to which the contractors agree. That system, after all, is the one that specifies which substantive considerations make systematic contributions to the moral status of action. Many of these considerations – paradigmatically those having to do with causing harm to others – will systematically count against actions. But considerations that involve the avoidance of harm to self will make systematic contributions to moral status as well. In particular, they will function to morally justify behavior that would otherwise be ruled out, morally, by other-regarding reasons. So such self-regarding considerations will count as moral reasons – though they will only provide moral justifications, not moral requirements. Now, when we redirect our attention to determining the *rational* status of an action, the substantive facts that constitute these self-regarding moral reasons will also constitute reasons: generic practical reasons. When we consider them as generic practical reasons, however, their contributions will be distinct from those they make when we consider them as moral reasons. In particular, one *is* rationally required to avoid harm to self, unless one has sufficient justification. But there is no *moral* requirement of the same sort. Determining the moral status of an act is simply a different thing from determining its rational status. This is true even when all the same considerations are relevant. That is why the moral status of an action can be different from its rational status.

Consider a certain fact about an action: that it will break someone else's legs. This fact makes a systematic contribution to the moral status of the action. It also makes a systematic contribution to the rational status of the action. Those contributions are, however, quite different. The fact that an action will break someone's legs seems, morally, to *require a refusal* to perform it, at least if there is no other consideration that would provide quite a strong moral justification for doing so (for example, that the act is required to save the life of the same person, or of the agent). But when we turn to consider the *rational* status of the act, the very same leg-breaking fact does not seem to play the same sort of role. Mafia enforcers are not *rationally* required to refuse to break people's legs – at least if they are moderately well-paid for that job. Still, the fact that an act will break someone's legs *is* relevant to the rational status of

34 Smith (2015), p. 192

the enforcer's choices. After all, this fact would rationally *justify* the enforcer in refusing to perform the action, and in attempting to get out of the enforcing business entirely – despite the well-known risks of such attempts.

I do not want to deny that it can seem quite natural and commonsensical to talk about moral reasons competing with non-moral reasons. Indeed, there is a way of interpreting such talk that is perfectly coherent. Suppose one is choosing between two options: (i) breaking someone's legs, as part of one's job as a Mafia enforcer, and (ii) refusing to do so. Since the moral relevance of the fact that an action will break someone's legs will often be very salient, it may be natural to think of it in a morally tinged way, and to describe it as 'a moral reason' that favors option (ii). And yet, as noted, that moral reason is simply a naturalistic fact about the action – a fact that also counts as a generic practical reason. On the other hand, a fact about (i) is that by choosing it, one avoids the risks that come with defying a Mafia boss. This fact is more salient as a generic practical reason than as a moral reason. Given all this, it would be natural, even when discussing the *rational* status of (ii), to describe it in terms of a conflict between a moral reason and a non-moral reason. Still, such a description does not presuppose the unitary view of reasons.

4. Rational Options

We are now in a position to say something about Smith's current embrace of a strong moral rationalism – a position that he once criticized as 'flawed in a quite decisive way'.[35] I think part of the explanation *could* be that he has lost sight of an option between moral rationalism and a position we might call moral *anti-rationalism*. Moral anti-rationalism is the view that rationality, at least sometimes, *requires* immoral action. Smith points out that Hume's view (or, in any case, one that is often attributed to Hume) is anti-rationalist in this way, since on such a view someone with sufficiently evil basic desires would be irrational to avoid causing pain to babies just for the fun of it. Smith wants, quite reasonably, to reject this as wildly implausible. But as part of his rejection of moral anti-rationalism, what he says is the following:

> Absent skepticism about both moral requirements and moral responsibility, it seems that moral requirements *must*, in some way, reduce to rational requirements.[36]

But the 'must' here is misplaced. Moral rationalism is not the only alternative to moral anti-rationalism. A middle position, which seems (to me) by far

35 Smith (2002), p. 121.

36 Smith (2013), p. 13, my italics. It is interesting that one of Smith's most comprehensive critics, Michael Bukoski, seems to share Smith's assumption. Bukoski (2016, p. 143) suggests that if 'we think that morality and rationality cannot conflict', then we would 'have good reason to identify moral requirements with rational requirements'.

the most plausible, is that moral requirements, while *not* always rational requirements (as against moral rationalism), *are* always rationally permissible (as against moral anti-rationalism). We can call this view *moral permissibilism*. As noted in the introduction, this is a view that Smith himself once explicitly endorsed. On such a view, those who act immorally need not be acting irrationally, though it always would have been rationally permissible for them to have chosen a morally acceptable option instead. I will not rehearse, again, the arguments I have offered for this sort of position. But I will close by describing the view of practical rationality on which those arguments depend, and by explaining why, in defending such a view, it helps to endorse the relational view of reasons, as against the unitary view.

According to the relational view, practical rationality is one normative domain among many, and the rational status of an action – its being rationally required, permitted, or prohibited – is determined by a function of the contributions of the relevant facts about that action. These facts, which we have been calling 'generic practical reasons', count as such reasons in virtue of their making systematic contributions to such rational statuses. These reasons are simply certain naturalistic matters of fact: that someone will be hurt in such-and-such a way; that the agent will be benefitted in such-and-such a way; that some third party will avoid being hurt in a way that they would otherwise be hurt, and so on. On the particular view I have defended, the content of practical rationality can be summarized, at least to a first approximation, by the following principle:

> P: It is rationally required to act so as to avoid harms to *oneself*, unless one has adequate justification for not avoiding those harms. Justification is provided by the prospect of avoiding other harms, or gaining benefits, for *anyone*.[37] Being justified entails being permissible, but does not entail being required.

Harms include such things as death, pain, disability and loss of freedom, while benefits include such things as pleasure, ability, and freedom. Principle P entails that *only* reasons having to do with avoiding harm for oneself have any rationally requiring strength. That is, it is only such reasons that one is *ever* rationally required to act on. Nor is one always required to act on them, since in many cases one will have adequate justification for acting against them. Principle P also gives a systematic role to altruistic reasons, despite their inability to require anything: they can provide the rational justification one needs for acting in ways that will bring harm to oneself.[38] What determines

37 See Gert (2007a), p. 544.

38 It is possible to modify the view so that altruistic generic practical reasons have some minimal requiring strength. This does not affect anything of significance for this paper. In particular, the view of rationality will still support moral permissibilism, rather than moral rationalism or moral anti-rationalism.

the *adequacy* of such a justification – that is, what determines the requiring and justifying strengths of generic practical reasons – is a complex matter which I discuss elsewhere.[39]

One important entailment of principle P is that if one would be rationally justified in suffering a certain harm because of some reason that has to do with one's own interests, then one would be equally justified in suffering that same harm because of a reason that has to do with anyone else's similar interests.[40] For example, if I would be justified in breaking my arm in order to save my *own* life, I would also be justified in breaking my arm in order to save *your* life, or a *stranger's* life. Again, I will not defend this account of practical rationality here. But I will note that it is extensionally very plausible. It allows altruistic sacrifices to count as rationally permissible, but not rationally required. It entails that willingly suffering harms is irrational if no one at all will receive any compensating benefits. And it is very much in line with an understanding of irrationality that captures the sort of behavior definitive of mental illnesses like depression or compulsions or phobias, as well as the self-destructive "acting out" that strong emotions can produce.

Armed with a notion of practical rationality of the sort just described, we can construct a moral theory. It might be a contractualist one, according to which morality is the system of sanctionable rules that would be put forward by rational agents under certain conditions. Again, these conditions might be a matter of being behind something like Rawls' veil of ignorance. Or they might include the stipulated possession of a desire to come to an agreement. Such a moral system will specify the sorts of considerations that place an action in need of moral justification in the first place. We can say that these considerations provide *pro tanto* moral prohibitions; one is morally prohibited from acting against these reasons unless one has sufficient justification. It is extremely plausible that rational contractors who meet the suggested "certain conditions" would put forward a system according to which one is *pro tanto* morally prohibited from killing anyone, or hurting them, or deceiving them, or breaking promises to them, and so on. This is true quite regardless of whether they themselves would be motivated to abide by the agreement when no longer under those "certain conditions". The system would also need to specify what sorts of considerations count as providing adequate justification for acting against such *pro tanto* moral requirements. Importantly, having adequate justification only entails that one is morally *permitted* to act against the *pro tanto* requirement. As in many other domains – the law, for example, and, I have argued, practical rationality – one is not morally *required* to perform an act simply because one has sufficient moral justification for performing it.

---

39 Gert (2012).

40 In Gert (2004, pp. 99–101) I call this feature 'the agent-neutrality of justification'.

Because – on the contractualist view we are supposing – *pro tanto* moral prohibitions and *pro tanto* moral justifications make systematic contributions to the overall *moral* status of an action, we can also call them *moral reasons*. As explained above, the justifying and requiring strengths that a certain consideration has *as a moral reason* might be quite different from the justifying and requiring strengths it has *as a generic practical reason*. The relevant functions – functions from act-descriptions to normative statuses – are simply distinct, and can take the same substantive facts into account in different ways. This makes room for the possibility that when an act is *morally required* because of some non-normative fact about it – for example, that it is the only way to avoid causing some third party a significant harm – that very same non-normative fact only provides a *rational justification* for performing the act, given that there is some degree of self-sacrifice involved. This allows us to explain why, even if one would be morally required to tell the truth and give someone an alibi for a crime for which they will otherwise be wrongly convicted, one is not *irrational* for lying in that scenario in order to save one's own skin. And it allows us to explain why one would not be irrational for telling the truth in such circumstances either. Moral permissibilism is consistent with the rational permissibility of both options. Moral rationalism is not. And moral anti-rationalism is implausible.

Is moral permissibilism reasonable as a general view? More precisely, is it reasonable to think that *whenever* morality requires some personal sacrifice, something about the action makes it rationally permissible to make that sacrifice? On a plausible contractualist moral theory, the answer seems clearly to be 'yes'. Rational contractors would not include, in the moral system they advocate, any requirements to perform actions that would be irrational. They desire, as we ourselves do, that other people behave as they are morally required to behave. But they, also like us, do not want anyone with whom they are concerned – principally including themselves – to behave irrationally. And they also know that no normal person will, knowingly, follow a code that requires them to behave in irrational ways. So the moral system they put forward – the output of the contractualist view of morality – will never require irrational behavior. In particular, the moral system they put forward will only require an action that involves a personal sacrifice if there are other features of the action that rationally justify it.

References

Berker, Selim. 2007. 'Particular Reasons'. *Ethics* 18(1), pp. 109-139.

Bukoski, Michael. 2016. 'A critique of Smith's constitutivism'. *Ethics* 127(1), pp. 116-146.

Copp, David. 1997. 'Belief, Reason, and Motivation: Michael Smith's The Moral Problem'. *Ethics* 108, pp. 33-54.

Copp, David. 2009. 'Toward a pluralist and teleological theory of normativity'. *Philosophical Issues 19, Metaethics*, pp. 21-37.

Enoch, David. 2020. 'Constitutivism: On rabbits, hats, and holy grails'. In R. Chang and K. Sylvan (Eds.), *The Routledge Handbook of Practical Reason*. Routledge, pp. 336-348.

Gert, Joshua. 2004. *Brute Rationality: Normativity and Human Action*. Cambridge: Cambridge University Press.

Gert, Joshua. 2007a. 'Normative Strength and the Balance of Reasons'. *Philosophical Review* 116(4), pp. 533-562.

Gert, Joshua. 2007b. 'Moral Reasons and Rational Status'. *Canadian Journal of Philosophy* Supp. Vol. 33, pp. 171-196.

Gert, Joshua. 2008. 'Michael Smith and the Rationality of Immoral Action'. *The Journal of Ethics* 12(1), pp. 1-23.

Gert, Joshua. 2016. 'The Distinction between Justifying and Requiring: Nothing to Fear'. In *Weighing Reasons*, B. Maguire and E. Lord (Eds.). Oxford: Oxford University Press, pp. 157-172.

Harman, Elizabeth. 2020. 'There is no moral ought and no prudential ought'. In *The Routledge Handbook of Practical Reason*, eds. R. Chang and K. Sylvan. New York: Routledge, pp. 438–456.

Horgan, Terry and Mark Timmons. 1996. 'Troubles for Michael Smith's Metaethical Rationalism'. *Philosophical Papers* 25, pp. 203–231.

Lenman, James. 1999. 'Michael Smith and the Daleks: reason, morality, and contingency'. *Utilitas* 11(2), pp. 164–177.

Noordhof, Paul. 1999. 'Moral Requirements are Still not Rational Requirements'. *Analysis* 59, pp. 127–136

Raz, Joseph. 1997. 'Incommensurability and Agency'. In R. Chang (Ed.), *Incommensurability, Incomparability, and Practical Reason*. Cambridge, MA: Harvard University Press, pp. 110–128.

Sadler, Brook. 2003. 'The Possibility of Amoralism: A Defense Against Internalism'. *Philosophy* 78, pp. 63–78

Sagdahl, Mathea Slåttholm. 2022. *Normative Pluralism: Resolving Conflicts Between Moral and Prudential Reasons*. Oxford: Oxford University Press.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Wiley-Blackwell.

Smith, Michael. 1996. 'Normative reasons and full rationality: reply to Swanton'. *Analysis* 56, pp. 160–68.

Smith, Michael. 1999. 'The Definition of "Moral"'. In D. Jamieson (Ed.), *Singer and His Critics*, Oxford: Blackwell, pp. 38–63.

Smith, Michael. 2002. 'Bernard Gert's Complex Hybrid Conception of Rationality'. In R. Audi and W. Sinnott-Armstrong (Eds.), *Rationality, Rules, and Ideals: Critical Essays on Bernard Gert's Moral Theory*. Lanham: Rowman and Littlefield Publishers, pp. 109–123.

Smith, Michael. 2010. 'Beyond the error theory'. In R. Joyce and S. Kirchin (Eds.), *A world without values: essays on John Mackie's moral error theory*. New York: Springer Science & Business Media, pp. 119–139.

Smith, Michael. 2011. 'Deontological Moral Obligations and Non-Welfarist Agent-Relative Values'. *Ratio* 24(4), pp. 351–363.

Smith, Michael. 2013. 'A constitutivist theory of reasons: Its promise and parts'. *Law, Ethics and Philosophy* 1, pp. 9–30.

Smith, Michael. 2015. 'The Magic of Constitutivism'. *American Philosophical Quarterly* 52, pp. 187–200.

Smith, Michael. 2018. 'Three kinds of moral rationalism'. In K. Jones and F. Schroeter (Eds), *The many moral rationalisms*. Oxford: Exford University Press, pp. 48–69.

Swanton, Christine. 1996. 'Is the Moral Problem Solved?'. *Analysis* 56, p. 156.

Wolf, Susan. 1982. 'Moral Saints'. *Journal of Philosophy*, pp. 419–439.

Caj Strandberg
University of Oslo
c.s.strandberg@ifikk.uio.no

Original Scientific Paper
UDC 17.021.1/.2
17.034:165.6
17 Смит М.
Received: August 28, 2024
Revised: October 01, 2024
Accepted: October 04, 2024

SMITH ON THE PRACTICALITY AND OBJECTIVITY OF MORAL JUDGMENTS

Abstract
The moral problem presented by Michael Smith in his seminal book with the same name consists of three claims that are intuitively plausible when considered separately, but seem incompatible when combined: moral judgments express beliefs about objective moral facts, moral judgments are practical in being motivational, and beliefs are unable to motivate by themselves. An essential aspect of Smith's solution to the moral problem is the contention that moral judgments are both motivating for rational agents and objective. In this paper, I take a close look at Smith's arguments by considering his characterizations of rationality understood as coherence between attitudes. It is suggested that on this understanding of rationality and coherence, it has not been clearly shown that moral judgments are both practical and objective.

Keywords: Michael Smith · the moral problem · moral objectivity · practical rationality · motivational internalism · coherence

1. Introduction

The publication of Michael Smith's *The Moral Problem* in 1994 was a main event in metaethics. At that time, I was a fresh student with a bourgeoning interest in metaethics and I can still recall the fascination and admiration to which the book gave raise among teachers and fellow students. Our response was confirmed by the deep impact of the book. The influence of Smith's work is evinced by other philosophers discussing issues it brings to attention and arguments it presents. What is more, it is confirmed the by the numerous critical comments his work generated and still generates. As often is the case in philosophy, the better a work is, the more it gets criticized. Further discussion of Smith's work is thus further evidence of the importance of his contribution.

In some respects, the metaethical debate looks different than it did in 1994. One key difference concerns how philosophers conceive of the relation between normative reasons and practical rationality. According to the approach represented by Smith, normative reasons should be explained in

 DOI: 10.5937/BPA2437059S

terms of rationality.[1] In brief, an agent has a normative reason to perform an action insofar as she would have a pro-attitude of some sort towards the action if she were practically rational. In recent times, it has been common to assume, conversely, that rationality should be explained in terms of responding to reasons.[2] In brief, an agent is practically rational insofar as she responds to the normative reasons of which she is aware.[3] This is not the place to evaluate the *pros* and *cons* of respective strategy. However, in one important respect I think the former approach has an important advantage: It has the potential of being more explanatory potent. On this strategy, it might be possible to explain reasons in terms of rationality and then explain rationality in terms of coherence between attitudes. Thus, one normative notion—reasons—is explained in terms of another normative notion—rationality—of which one provides a substantive account by employing the notion of coherence. By contrast, on the latter strategy rationality is explained in terms of normative reasons of which it seems difficult to say anything relevantly informative.[4] Thus, one normative notion—rationality—is explained in terms of a more fundamental normative notion—reasons—and not much substantive can be said about the latter. The present paper can be seen as a discussion of whether Smith's particular version of the former strategy turns out to be successful.

2. The Moral Problem

The moral problem consists in the fact that three claims that, considered separately, are intuitively plausible seem incompatible when combined with one another.

The first claim states that moral judgments are objective:

> (**1**) *Objectivity of Moral Judgments*: 'Moral judgements of the form "It is [morally] right that I φ" express a subject's beliefs about an objective matter of fact, a fact about what it is right for her to do.'[5]

According to Smith's view of moral objectivity, (**1**) entails that (**i**) moral judgments consist in beliefs that have truth value and (**ii**) there are non-arbitrary criteria of what count as relevant considerations of whether an

1 See e.g. Williams (1981), pp. 101–113; Korsgaard (1996), pp. 188–221, and Markovits (2014), Ch. 3. In Strandberg (2024), pp 256–273, I put forward a Neo-Humean view that represents a version of this approach. In what follows, 'rationality' refers to practical rationality.

2 See e.g. Parfit (2011), Ch. 1; Kiesewetter (2017), Ch. 7, and Lord (2018), Ch. 1–2.

3 The shift in the debate has been motivated by considerations as regards the connections between reasons, rationality, and different interpretations of rational requirements. See e.g. Broome (2013), Ch. 7–10 and Kolodny (2005), pp. 509–563.

4 However, see e.g. Gert (2004), Ch. 7 and Schroeder (2007), Ch. 4.

5 Smith (1994), p. 12. Concise accounts of Smith's view are found in Smith (1996a), pp. 277–302 and Smith (1998), pp. 149–172. See also Smith (1989), pp. 89–174.

action is right, where these criteria are set solely by the circumstances of situations. As Smith puts it, 'the only relevant determinant of the rightness of an act is the circumstances in which the action takes place. If agents in the same circumstances act in the same way then either they both act rightly or they both act wrongly'.[6]

The second claim states that there is a conceptually necessary connection between moral judgments and motivation, a view usually referred to as 'motivational internalism':[7]

> (**2**) *Practicality of Moral Judgments*: 'If someone judges that it is [morally] right that she φs then, *ceteris paribus*, she is motivated to φ.'[8]

However, Smith recognizes that an agent might judge that an action is right for her to perform but still not be motivated to perform it if she suffers from some form of practical irrationality, such as weakness of will, depression, apathy, addiction or compulsion. He therefore thinks that (2) should be reformulated in the following way:

> 'If an agent judges that it is [morally] right for her to φ in circumstances C, then either she is motivated to φ in C or she is *practically irrational*.'[9]

The third claim states a standard view of motivation:

> (**3**) *Humean Theory of Motivation*: 'An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume's terms, distinct existences.'[10]

These three claims create the moral problem: According to (**1**), moral judgments express beliefs about objective moral facts. According to (**2**), moral judgments are necessarily accompanied by motivation in rational agents. According to (**3**), beliefs and desires are distinct existences such that beliefs cannot motivate to action but only desires can do so. Thus, (**1**) and (**2**) entail that there is a necessary connection between beliefs about objective moral facts and desires. However, according to (**3**) this cannot be the case.

---

6 Smith (1994), p. 5.

7 For a discussion of different types of motivational internalism, see Björklund et al. (2012), pp. 123–137. In Strandberg (2011), pp. 341–369 and Strandberg (2012), pp. 87–122, I propose a version of externalism that employs the notion of generalized conversational implicature to explain the close connection between moral utterances and motivation. In Strandberg (2013), pp. 25–51, I argue that Smith's version of internalism faces a dilemma that can be solved by this version of externalism. Smith's main argument against externalism and for internalism is the much discussed fetishist argument (Smith (1994), pp. 71–76 and Smith (1997), pp. 11–117). In Strandberg (2007), pp. 249–260, I defend externalism from this argument and provide an explanation of moral motivation.

8 Smith (1994), p. 12.

9 Smith (1994), p. 61. Emphasis added.

10 Smith (1994), p. 12. For a critical discussion of the Humean theory of motivation, see Arruda (2019), pp. 157–178.

As can be seen, Smith talks about *the* moral problem. However, strictly speaking, I think there are *two* moral problems. The first concerns the connection between (**1**) and (**2**). The problem is that if (**1**) moral judgments consist in beliefs about objective moral facts, it is difficult to see how (**2**) moral judgments could be necessarily accompanied by motivation. The second concerns the relation between (**1**), (**2**), and (**3**). The problem is that if (**1**) moral judgments consist in beliefs and (**2**) moral judgments are necessarily accompanied by motivation, it follows that there is a necessary connection between beliefs and desires, which seems incompatible with (**3**). In order not to complicate matters, I will not distinguish between the two moral problems in what follows.

3. A Simple Solution to the Moral Problem?

It might be suggested that there is a simple manner to solve the moral problem. Consider the first characterization of (**2**), the Practicality of Moral Judgments, according to which it is conceptually necessary that if an agent judges that it is right that she φs, then she is, *ceteris paribus*, motivated to φ. That is, on (**2**) moral judgments are by conceptual necessity *accompanied* by motivation. Now, this can be explained in different manners. According to one alternative, (**2**) is explained by it being conceptually necessary that a moral judgment partly or wholly *consists* in a motivational state. In the literature, this view is often formulated by saying that motivation is 'internal' or 'intrinsic' to moral judgments. According to another alternative, (**2**) is explained by the fact that we *classify* an agent's judgment as a *moral* judgment only if she is accordingly motivated. This alternative is compatible with the nature of the judgment itself not playing any part in the explanation of why the agent is motivated. In the literature, it is standardly presumed that (**2**) should be understood in accordance with the first alternative, and it is only fairly recently that it has been noticed that it is compatible with the second alternative.[11]

We might now return to the moral problem. The problem is assumed to consist in the following: (**1**) and (**2**) entail that there is a necessary connection between objective moral beliefs and desires, but according to (**3**) this cannot be the case since beliefs are unable to motivate and only desires have this capacity. However, this contention presumes that (**2**) is read in accordance with the first alternative. It is only if (**2**) is understood to entail that a moral judgment wholly or partly consists in a motivational state that it together with (**3**) entails that moral judgments cannot consist in beliefs. However, if (**2**) is read in accordance with the second alternative, (**1**), (**2**), and (**3**) are fully compatible. According to this alternative, we classify a judgment as a moral judgment only if it is accompanied by motivation in the form of desires, which is fully compatible with moral judgments consisting in belies about objective moral facts.

11 See Tresan (2006), pp. 143–165. See also e.g. Strandberg (2011), pp. 342–347 and Francén (2020), pp. 366–379.

However, I do not think that the moral problem is so easily solved. The main reason is that the latter alternative is not a plausible explanation of Smith's version of (**2**). Recall that Smith, very plausibly, argues that (**2**) should be restated to say that it is conceptually necessary that if an agent judges that it is right that she φs, then she is motivated to φ *on condition that she is rational*. It seems reasonable to think that the fact that we classify an agent's judgment as a moral judgment only if she is accordingly motivated need not be explained by anything about the content of the judgment. It is merely a matter of how we choose to classify certain judgments. By contrast, it is much less plausible to maintain that the fact that we classify an agent's judgment as a moral judgment only if she is accordingly motivated *on condition that she is rational* has nothing to do with its content. It seems that the judgment needs to have a particular content which explains that an agent who makes it is guaranteed to be accordingly motivated only if she is rational. As we shall, see Smith provides such an account.

4. Smith's Solution to the Moral Problem

The solution Smith propose of the moral problem goes through three steps to be explained below. The general idea is this: Analyze rightness in terms of normative reasons. Analyze normative reasons in terms of what an agent would desire if she were rational. As part of this step, maintain that an agent is rational insofar as she has coherent desires and maintain that rational agents would 'converge' in their desires. Finally, analyze judgments about normative reasons as beliefs about what the agent would desire if she were rational in this sense. In this manner, it is thought that moral judgments can be shown to be both practical and objective at the same time as it is insisted that only desires can motivate.

The first step consists in a conceptually necessary claim about the connection between moral rightness and normative reasons:

> (**R**) *Rationalism*: 'If it is [morally] right for agents to φ in circumstances C, then there is a [normative] reason for those agents to φ in C.'[12]

The second step amounts to a platitude—which in Smith's vocabulary amounts to a conceptually necessary claim—about normative reasons according to which they consist in facts about what an agent would desire under full rationality:

> (**P**) *Platitude about Reasons*: 'What we have normative reason to do is what we would desire that we do if we were fully rational'.[13]

12 Smith (1994), p. 62.

13 Smith (1994), p. 150. Smith sometimes uses 'reason' in the singular without clarifying whether he has in mind a *pro tanto* reason or *pro toto* reason (all-things-considered strongest reason). Similarly, he sometimes uses 'desire' without clarifying whether he has

After having presented this connection between normative reasons and rational desiring, Smith offers an account of practical rationality. A fully rational agent (**i**) has no false beliefs; (**ii**) has all relevant true beliefs, and (**iii**) has the capacity to deliberate correctly.[14] The last condition entails that the agent is able to imagine relevant facts and to desire in accordance with means-ends norms. Most importantly, an agent who has deliberated correctly has a set of desires that is 'systematically justified', which means that it is completely coherent and unified.

According to Smith, our concept of normative reasons entails that fully rational agents would converge in their desires about what to do. In conjunction with (**P**) this claim entails that all agents, given sameness in circumstances, have the same normative reasons, that is, normative reasons with the same content. Indeed, Smith maintains that if reasons were thought to be relative to actual desires, they would not be normative since they would be arbitrary, and error-theory about normative judgments would follow.[15] In view of the fact that normative reasons have this particular nature, they can be said to be *categorical*: They are not relative to actual desires, and agents who find themselves in the same circumstances have the same normative reasons.

The last step in Smith's solution of the moral problem is to argue that there is a conceptually necessary implication from beliefs about normative reasons to rational desiring:

> (**I**) *Implication from Normative Beliefs to Rational Desiring*: 'If an agent believes that she has a normative reason to φ, then she rationally should desire to φ'.[16]

---

in mind *a* desire or strongest desires. At the same time, it seems to be assumed that the strength of reasons corresponds to strength of desires in rational agents. It is plausible to think that these unclarities and assumptions complicate his argumentation. For a discussion of Smith's view of the strength of reasons and desires, and of the relevance of the distinction between the rationally requiring and rationally permissive strength of reasons, see Gert (2008), pp. 1–23.

14 Smith's conception of full rationality is a modified version of Bernard Williams's account of this notion. See Williams (1981), pp. 101–105.

15 '[T]o suppose that our concept of what constitutes a rational justification could be radically relative in some way to the interests or desires of those who make claims about what is rationally justified, so that the considerations that rationally justify relative to that agent may fail to justify relative to another—is to suppose, quite incoherently, that something completely arbitrary—the mere fact that a particular agent who is making a claim about rational justification happens to have the contingent interests or desires that she happens to have—could in some way constitute a normative fact: a fact about rational justification. But this is incoherent because the only decisive point we can make about normativity is that arbitrariness, as such, always undermines normativity' (Smith (1997), p. 90).

16 Smith (1994), p. 148. As Geoffrey Sayre-McCord argues, the implication should be read: 'If an agent believes she has reason to φ then if she is rational she will desire to φ' (Sayre-McCord (1997), p. 64).

In Smith's view, (**P**) entails (**I**) on the assumption that an agent believes that she has a normative reason to φ. He argues in the following way. If an agent believes that she has a normative reason to φ, it follows from (**P**) that she believes that she would desire to φ if she were fully rational. Accordingly, if the agent believes that she has a normative reason to φ, but does not desire to φ, she is irrational because she believes that she would desire that she φs if she were fully rational, but nevertheless does not desire to φ. The agent's irrationality is said to consist in a special kind of incoherence between her attitudes:

> Imagine two agents who believe that they would desire that they φ in C if they had a maximally informed, coherent and unified set of desires. One of these agents also desires that she φs in C, but the other does not desire that he φs in C. What can we say about the relative merits of these two agents' psychologies, straight away, given just what we've said? The obvious thing to say, it seems to me, is that the former psychology exhibits more in the way of coherence than the latter. *The latter agent fails to have a desire that he believes he would have if he had a maximally coherent set of desires, and this fact, all by itself, constitutes a kind of incoherence or disequilibrium in his overall psychological state*. The former agent's desires do not suffer any such disequilibrium or incoherence. Even if her belief is false, she still enjoys a sort of coherence, or equilibrium, simply in virtue of the fact that she has a matching desire to φ in C.[17]

We have thus far been concerned with normative reasons in general rather than moral reasons in particular.[18] What makes normative reasons to *moral* reasons, what characterises this subset of normative reasons, is that they are demarcated by platitudes concerning the content of morality.[19] These platitudes provide together with (**R**) and (**P**) Smith's analysis of moral rightness:

> *Moral Rightness*: 'φ-ing in circumstances C is [morally] right if and only if we would desire that we φ in C, if we were fully rational, *where φ-ing in C is an act of the appropriate substantive kind*'.[20]

We can now see how Smith aims to solve the moral problem.

First, (**R**) Rationalism and (**P**) the Platitude about Reasons are maintained to entail (**1**) the Objectivity of Moral Judgments. From (**R**) and (**P**), understood

---

17 Smith (1995b), p. 166. Emphasis added. Cf. Smith (199a5), pp. 126–127 and Smith (1994), pp. 177–179.

18 This has caused some misunderstanding of Smith's view; see Swanton (1996), pp. 155–160 and Smith (1996b), pp. 160–168.

19 Smith (1994), pp. 40–41, 183–184.

20 Smith (1994), p. 184.

in accordance with Smith's conception of full rationality, it follows that whether it is right that an agent φs is determined by considerations that are not relative to her actual desires but determined by objective circumstances. That is, facts about what is morally right, a subclass of facts about normative reasons, are constituted by categorical requirements of rationality. Thus, (**1**) follows.

Second, (**R**) Rationalism and (**P**) the Platitude about Reasons are maintained to entail (**2**), the Practicality of Moral Judgments. We have already seen that Smith argues that (**P**) entails (**I**), the Implication from Normative Beliefs to Rational Desiring, on the assumption that an agent believes that she has a normative reason to φ. Now, from (**R**) and (**P**), together with this assumption, follows a moral analogue to (**I**): An agent who judges that it is morally right to φ—i.e. believes that she has a moral reason to φ—would be motivated to φ on condition that she is rational. Thus, (**2**) follows.

Lastly, the three steps—(**R**) Rationalism, (**P**) the Platitude about Reasons, and (**I**) the Implication from Normative Beliefs to Rational Desiring—are maintained to be compatible with (**3**) the Humean Theory of Motivation, the view that beliefs cannot motivate and that only desires are able to fulfil this function. Importantly, it is the claim that an agent's moral judgment consists in a belief about what she would desire if she were fully rational which explains that it is accompanied by a desire—motivation—to perform the action given that she is rational. Is not claimed that the moral belief in itself is motivating in insolation from any desire. Thus, the moral problem is solved.

In this paper, I will mainly be concerned with Smith's contention that moral judgments are both practical and objective. More precisely, I will argue that Smith's conception of (**P**), the Platitude about Reasons, is such that (**2**) the Practicality of Moral Judgments and (**1**) the Objectivity of Moral Reasons have not clearly been shown to follow. As a result, it might be doubted whether Smith has succeeded to solve the moral problem. Smith has developed his metaethical views after *The Moral Problem*.[21] In what follows, I will primarily focus on Smith's view as it is formulated in the book, since the present issue of *Belgrade Philosophical Annual* is dedicated to the thirtieth anniversary of its publication.

5. Example Model and Advice Model

As was seen in the last section, (**P**) the Platitude about Reasons is a vital step in Smith's solution to the moral problem. We might start with considering a certain ambiguity in (**P**) identified by Smith in his seminal paper from 1995.[22]

---

21 See in particular Smith (2009), pp. 98–125 and Smith (2012), pp. 309–331. In understanding Smith's view, I have learned a lot from David Brink's, David Copp's, and Geoffrey Sayre-McCord's comments on *The Moral Problem*: Brink (1997), pp. 4–32; Copp (1997), pp. 33–54, and Sayre-McCord (1997), pp. 55–83.

22 Smith (1995a), pp. 110–112, 125–129. For criticism of Smith's paper, see Johnson (1997), pp. 619–625.

The simplest reading of (**P**) is:

> (**P.1**) *Example Model*: A (less than fully rational) agent has a normative reason to φ in a given situation if and only if she would desire to φ in that situation if she were fully rational.

According to (**P.1**), what an agent's fully rational self would desire to do forms an *example* to her less than fully rational self. However, this is not how (**P**) should be understood in Smith's view. What we have normative reasons to do, which Smith equates with what is *desirable* that we do, is not necessarily what we would desire to do if we were fully rational. As we actually are, we have beliefs and desires others than those we would have if we were fully rational. Accordingly, we might, as fully rational agents, be motivated to perform actions that we, as we actually are, have no reasons to perform. Conversely, we might, as we actually are, have reasons to perform actions that we would not, as fully rational agents, be motivated to perform. Smith has a well-known example that is designed to show why this is the case:[23]

> *Squash Example*: You have been defeated in a game of squash in a way that you find humiliating. If you were fully rational, you would have been calm and walked to your opponent to shake hands with him. However, as you are not fully rational, you are consumed by a desire to smash your opponent in the face with your racket.

In Smith's view, what it is desirable that you do is not to stride over and shake hands with your opponent, but to leave the court as soon as possible.

In Smith's view, (**P**) should instead be read as follows:

> (**P.2**). *Advice Model*: A (less than fully rational) agent has a normative reason to φ in a given situation if and only if her fully rational self would desire that her less than fully rational self φs in that situation.

According to (**P.2**), an agent's fully rational self *advises* her less than fully rational self about what to do. In the squash example, leaving the court is not what your fully rational self would desire to do, since she does not have a desire to smash your opponent in the face. However, walking away is what your fully rational self would desire that you, being less than fully rational, do, since it is the preferable alternative given your desire.

6. Doubt about the Advice Model

We might start with considering Smith's motivation for rejecting (**P.1**) the Example Model in favour of (**P.2**) the Advice Model. An intuitive objection against Smith's argument in relation to the squash example might be that it *is*

23 Smith (1995a), pp. 111–112.

desirable that you shake hands with your opponent and that this is the case quite independently of your desires.

One possible reply is that it is desirable that you do not walk to your opponent to shake hands with him but leave the court because it would have the most favourable consequences.[24] If you stride up to your opponent to shake hands, you will smash him in the face, but if you walk away that will not happen. However, it might be worried that the reply is not satisfactory. First, it seems unclear whether Smith is in the position to appeal to favourable consequences of actions without having a substantive view of rationality in place.[25] Second, it might be objected that what is desirable that you do is to shake hands with your opponent *without* smashing him in the face and that *this* would have the most favourable consequences. The sad fact that you will smash your opponent if you walk to him to shake hands is neither here nor there for what is desirable that you do or what has preferable results.

It might perhaps be protested that the objection betrays a misunderstanding of the example. One possible interpretation is the following. What is wrong with (**P.1**) is that it implies that what is desirable is that you do something that you could do if you were fully rational, but that you *cannot* do because you have desires you would not have if you were fully rational. That is, (**P.1**) conflicts with an analogue to 'ought implies can': It is desirable that an agent performs a certain action only if she could get herself to perform it.[26]

In view of the extensive debate on 'ought implies can', it would be premature to assume any particular view about the dictum 'desirable implies can'. However, it is plausible to think that you in the relevant sense *can* walk to your opponent and shake hands with him without smashing him in the face because you would do so *if* you desire to and your desire to smash your opponent were not allowed to prevail. This suggestion gains support from the common idea that an agent could have performed a certain action insofar as she would have performed it if she had desired or willed to. Moreover, the fact that an agent has or lacks a certain desire is ordinarily not considered as an acceptable excuse for her performing an undesirable action or not performing a desirable action, in contrast to other excuses, such as physical obstacles or ignorance.

However, let us grant that Smith's argument against (**P.1**) and in favour of (**P.2**) applies in cases where (**P.1**) would entail that it is desirable than an agent performs an action she is unable to perform. Nevertheless, I think there is an argument against (**P.2**) that applies in other cases. We can think of cases where (**P.2**) entails that it is desirable that an agent performs a certain action,

24 Cf. Smith (1995a), p. 111.

25 I owe this point to Joshua Gert.

26 Cf. Johnson (1997), pp. 619–625.

but where it is desirable that she performs quite another action that she is able to perform. Consider a slightly modified version of the squash example:

> *Squash Example Modified*: You are not such that you cannot avoid smashing your opponent in the face if you walk to him to shake hands. However, you have a strong desire to do so. In consideration of this desire, it is very probable that you will smash him in the face if you stride up to him to shake hands.

In view of your fully rational self's favourable epistemic position and her resulting information about probabilities, it seems reasonable to think that she would not desire that you stride up to your opponent to shake hands, but that you walk away. Moreover, in view of your fully rational self's epistemic position and information about probabilities, it seems that she would desire this even if, as things ultimately turn out, you walk to your opponent and shake hands with him without smashing him in the face. However, in this case it seems less plausible to maintain that what is desirable that you do is to walk away instead of striding up to your opponent to shake hands.

In the type of case under consideration, (**P.2**) does not get support from the 'desirable implies can' principle. We are considering cases with the following features: (**i**) An agent's fully rational self would desire that her less than fully rational self performs a certain action since the latter suffers from an imperfection as a result of lacking rationality; (**ii**) If her less than fully rational self did not suffer from this imperfection, her fully rational self would desire that her less than fully rational self performs quite another action, and **(iii**) Her less than rational self is able to perform another action. In this type of cases, (**P.2**) seems to give the wrong verdict. The reason is that what an agent's fully rational self would desire that her less than fully rational self do might depend on her less than fully rational self's actual desires at the same time as the latter is able to perform another action which we consider desirable.

It might be replied that your fully rational agent not only would predict that you would smash your opponent in the face. As she has no false beliefs and all relevant true beliefs, she would actually *know* what you would do. However, if we think of your fully rational self's epistemic position in this manner, it does not seem evident that she would have any desire concerning your actions. First, it might be asked whether your fully rational self would have any relevant desire concerning your actions if she already knows what you will do.[27] Smith is not explicit about whether knowledge of the future is included in a fully rational agent's 'all relevant true beliefs'. However, since it implies that a fully rational agent's desires depend on her beliefs, there is

27 This understanding of a fully rational agent's epistemic position would also be a problem for (**P.1**) the Example Model. Would you have a relevant desire concerning your actions if you already know what you will do?

some reason to think that knowledge of the future is not included.[28] Second, on Smith's view of desires in terms of 'direction of fit', desires essentially seem to be something that are directed towards an unknown future. According to this view, what characterizes a desire that *p* in contrast to a belief that *p* is their having different relations to the perception that not *p*. A belief that *p* is such that it would 'tend to go out of existence in the presence of a perception with the content that not *p*, whereas a desire that *p* tends to endure, disposing the subject in that state to bring it about that *p*'.[29] If an agent knows that the world fits her desire that *p* or that it will fit her desire that *p*, it is plausible to hypothesize that her desire that *p* will vanish, corresponding to a belief that *p* which will go out of existence in the presence of a perception that not *p*.

It might be replied that an agent who desires that *p* and who comes to know that *p* will be the case still will continue to desire that *p* because she is aware that *p* will not realize unless she desires that *p*. That is, she is aware that her having a desire that *p* is causally necessary for *p* to realize. However, this is not the situation of an agent's fully rational self. An agent's fully rational self has a desire *about* what her less than fully rational self is to do. Thus, it is not an agent's fully rational self having a desire that is causally necessary for something to realize; it is her less than fully rational self having a desire which has this function.

7. The Example Model and the Implication from Normative Beliefs to Rational Desiring

We saw above that Smith argues as follows in order to establish (**2**) the Practicality of Moral Judgments: (**P**) the Platitude about Reasons entail (**I**) the Implication from Normative Beliefs to Rational Desiring on the assumption that an agent believes that she has a normative reason to perform an action. From (**R**) Rationalism and (**P**) together with this assumption follows a moral analogue to (**I**): An agent who judges that an action is right for her to perform—i.e. believes that she has a moral reason to perform it—is motivated—desires—to perform the action on condition that she is rational. Thus, (**2**) has been established. This line of argument means that insofar as it turns out that (**P**) entails implausible versions of (**I**), there is reason to doubt that (**2**) has been established.

In the last section, it was found that there are reasons to doubt whether Smith has shown that (**P.2**) the Advice Model is to be preferred to (**P.1**) the Example Model. We might then return to (**P.1**). However, I think it might be argued that (**P.1**) entails implausible versions of (**I**).

If we insert the account (**P.1**) provides of normative reasons in (**I**), this implication should be read in the following way:

28 See Smith (1994), pp. 156–157 and Smith (1995a), pp. 112–113.

29 Smith (1994), p. 115.

(**I.1**) If an agent believes that she would desire to φ if she were fully rational, then she would desire to φ if she were rational.

In considering (**I.1**), it can be seen that 'rational' in the consequent cannot plausibly be understood in exactly the same way as 'fully rational' in the antecedent. In that case, the consequent would stipulate a condition of the truth of the implication that might be false according to the antecedent, namely that the agent is fully rational. We should then consider the possibility that 'rational' in the consequent concerns a relevant part of being 'fully rational'.

According to Smith's conception of rationality, it follows that an important aspect of an agent being fully rational is that she has a completely coherent set of desires. In the long quotation from Smith above, we saw that his argument that (**P**) entails (**I**) appeals to a particular type of incoherence between attitudes.[30] In line with Smith's argument, it might then be maintained that an agent is rational only insofar as she desires to do what she believes that she would desire to do if she had a completely coherent set of desires. We get the following version of (**I**):

(**I.2**): If an agent believes that she would desire to φ if her set of desires were completely coherent, then she would desire to φ if she desires to do what she believes she would desire to do if her set of desires were completely coherent.

However, explicated in this way it is evident that the conception of rationality referred to in the consequent (the second string of words underlined) is not part of the conception of rationality referred to in the antecedent (the first string of words underlined). The kind of rationality referred to in the antecedent concerns a desire an agent has given that it is coherent with her other *desires*. The kind of rationality referred to in the consequent concerns a desire an agent has given that it is coherent with a *belief* about what desire she would have if her set of desires were coherent.

The fact that the antecedent and the consequent refer to different conceptions of rationality gives rise to two difficulties. First, as far as I see, we have not been offered a clear account of the contention that coherence between beliefs and desires is part of rationality or, for that matter, that beliefs can cohere with desires. The latter is perhaps also something that might be questioned on the view that desires and beliefs have different 'directions of fit'.[31] Second, it seems that (**I.2**) involves a certain type of inconsistency. The conception of rationality referred to in the antecedent states that an agent is fully rational only if she has the desires she would have if her set of desires were

30 Smith (1995b), p. 166.

31 Sayre–McCord has raised a similar question. As far as I understand, Smith's response is more concerned with the coherence between desires than the coherence between beliefs and desires. See Sayre-McCord (1997), pp. 74–76 and Smith (1997), pp. 92–99.

completely coherent. However, the conception of rationality referred to in the consequent is in potential conflict with this notion. It states that an agent is rational only insofar as she has the desires that she believes she would have if she had a coherent set of desires. As a result, (**I**.**2**) involves two conceptions of rationality that make potentially conflicting demands: 'Desire as you would if you had a completely coherent set of desires' and 'Desire as you believe you would if you had a completely coherent set of desires'[32] The two demands will conflict with one another in cases where an agent has false beliefs about what she would desire if she had a completely coherent set of desires. Assume that an agent falsely believes that she would have a certain desire if she had a completely coherent set of desires. Assume further that she as a consequence of this belief comes to acquire the desire. The agent now has a desire which she rationally should have in accordance with the conception of rationality referred to in the consequent and she has thus become more rational on this conception. However, according to the conception of rationality referred to in the antecedent, it might be that she has become *less* rational because now she has a desire that she would not have if her set of desires were completely coherent.

8. The Advice Model and the Implication from Normative Beliefs to Rational Desiring

As we have seen, Smith's main motivation for preferring (**P**.**2**) the Advice Model to (**P**.**1**) the Example Model seems to be that the former squares better with our intuitive conception of what is desirable. However, it was found that there are reasons to doubt this assumption. In the last section, it was argued that (**P**.**1**) entails implausible versions of (**I**) the Implication from Normative Beliefs to Rational Desiring. It might then be suspected that the motivation for preferring (**P**.**2**) to (**P**.**1**) rather should be found in the conviction behind (**I**). We might therefore return to (**P**.**2**) and consider its relation to (**I**). However, as we shall see, it might be argued that also (**P**.**2**) entails an implausible version of (**I**).

In a central passage in the paper from 1995, Smith argues that the relation between beliefs about normative reasons and rational desiring supports (**P**.**2**) in favour of (**P**.**1**):

> Suppose, for instance, that you believe your fully rational self would desire to φ in the circumstances she faces; that this is the example she would set for you in her own word. Why should this have any effect at all on what you desire to do in the circumstances you face? [---] Coherence and unity do not argue in favour of acquiring a desire like hers because her example—marvelous though it is in the circumstances

32 Cf. Persson (1995), p. 151; Dancy (1996), p. 178, and Sayre-McCord (1997), pp. 55–83.

> in which *she* finds *herself*—doesn't engage with the circumstances in which *you* find *yourself*. This is not the case if instead we interpret the requirement in terms of the Advice Model. For then what you have to *believe is that your fully rational self would want your less than fully rational self to φ in the circumstances your less than fully rational self actually faces*. Your fully rational self's advice engages with your predicament because it is precisely tailored to it. You may still say 'So what?', of course, but if you do you simply reveal that you are unable to accept good advice; you reveal the extent to which your psychology fails in terms of norms of coherence and unity that define a systematically justified psychology. You thus simply betray your own irrationality.[33]

Thus, in effect the motivation for preferring (**P**.**2**) to (**P**.**1**) seems to come from an argument to the effect that (**P**.**2**) squares better with (**I**). In passing, it might be reflected that the contention seems troublesome in the context of Smith's overall argumentation. According to this line of argument, (**P**), together with the assumption that an agent believes that she has a normative reason, entails (**I**). However, if (**I**) is established because it follows from (**P**), it is questionable whether (**I**) can be used to support (**P**) or a particular version of it.

Nevertheless, I think it might be worried that (**P**.**2**) has essentially the same difficulty as (**P**.**1**): It entails an implausible version of (**I**). However, now things become a bit more complicated. According to (**P**.**2**.), it follows that an essential aspect of an agent being fully rational is that she, having a completely coherent set of desires, would desire that she, as she actually is, performs certain actions. In line with Smith's argument in the quotation above, it might then be maintained that an agent is rational only insofar as she desires to do what she believes that she, had she a completely coherent set of desires, would desire that she, as she actually is, do. The following convoluted version of (**I**) is the result:

> (**I**.**3**) If an agent believes that <u>she, had she a completely coherent set of desires, would desire that she, as she actually is, φs</u>, then <u>she would desire to φ</u> if she desires to do what she believes that she, had she a completely coherent set of desires, would desire that she, as she actually is, do.

However, it is doubtful whether (**I**.**3**) involves the relevant type of coherence. Consider the desire referred to in the antecedent (the first string of words underlined) and in the consequent (the second string of words underlined), respectively. In the first case, there is a desire which is the object of an agent's belief about what she would desire (if she had a completely coherent set of desires) that she do (as she actually is). That is, the desire which is the object

33 Smith (1995a), pp. 128–129. Emphasis added.

of the belief is an agent's hypothetical desire that she performs an action in a world where she is different from how she is in the world in which she has that hypothetical desire. In the second case, there is an agent's desire to perform an action as she is when she has the desire, in the actual world. The first kind of desire is a desire with the content 'that I do action x as I would be in that world'. The second kind of desire is a desire with the content 'to do action x as I am in this world'. The difference is indicated by the phrases 'desire that' and 'desire to'. In view of the fact that the contents of these desires are structured in different ways, it might be doubted that it is a matter of coherence of the requisite kind.

It should also be noticed that (**I.3**) is subject to the same type of problems as (**I.2**) noticed above. First, (**I.3**) refers to a conception of rationality that rests on the contention that desires can cohere with beliefs. Second, (**I.3**) involves two potentially conflicting demands of rationality: 'Desire as you, had you a completely coherent set of desires, would desire that you, as you actually are, do' and 'Desire as you believe that you, had you a completely coherent set of desires, would desire that you, as you actually are, do'. The two demands will conflict with one another in cases where an agent has false beliefs about what she, had she a completely coherent set of desires, would desire that she, as she actually is, do.

9. The Implication from Normative Beliefs to Rational Desiring and the Conception of Rationality

In the two proceeding sections, I have tried to argue that both versions of (**P**) the Platitude about Reasons—(**P.1**) the Example Model and (**P.2**) the Advice Model—entail implausible versions of (**I**) the Implication from Normative Beliefs to Rational Desiring. In view of the fact that (**2**) the Practicality of Moral Judgments is supposed to constitute a moral analogue to (**I**), there is reason to doubt that (**2**) has been established.

However, there might be a more direct manner of casting doubt on whether Smith has established (**2**). The reason is that (**2**), as originally conceived, and (**I**) seem to refer to different conceptions of rationality.

According to (**2**), if an agent judges that an action is right for her to perform then she is motivated to perform it on condition that she is rational. According to Smith's original conception of rationality, an agent is fully rational insofar as she has no false beliefs, all relevant true beliefs, is able to imagine relevant facts, and has a set of desires that is completely coherent and unified. The original conception of rationality is modelled to explain that an agent who judges that an action is right for her to perform might not be motivated to perform it if she suffers from common forms of irrationality, such as weakness of will or depression. For example, a depressed agent

might judge that an action is right for her to perform but nevertheless not be motivated to perform it in case she lacks some relevant true beliefs, cannot correctly imagine alternative actions available to her, or some of her desires are incoherent.

In the discussion above, we have found that the different interpretations of (**P**) entail different versions of (**I**). However, what is important in the present context is that the conception of rationality referred to in these versions of (**I**) differs from Smith's original conception of rationality as referred to in (**2**). Thus, in (**I.3**) the consequent entails that an agent is rational only insofar as she desires to do what she believes that she, had she a completely coherent set of desires, would desire that she, as she actually is, do.[34]

The fact that (**2**) and (**I**) refer to different conceptions of rationality has implications for whether Smith has succeeded to establish (**2**).

First, an agent fulfilling the modified conception of rationality is compatible with her not desiring to do what she judges is right for her to do. According to (**2**), if an agent judges that an action is right for her to perform—i.e. believes she has moral reason to perform it—then she is motivated—desires—to perform the action on condition that she is rational. In Smith's view, an agent has a normative reason to perform an action if she would desire that she performs it if she were fully rational. As we have seen, according to the original conception of rationality an agent is fully rational insofar as she has no false beliefs, all relevant true beliefs, is able to imagine relevant facts, and has a coherent and unified set of desires. By contrast, according to the modified conception of rationality referred to in the consequent of (**I.3**) an agent is rational only insofar as she desires to do what she believes that she, having a fully coherent set of desires, would desire that she, as she actually is, do. An agent being rational on the modified conception is evidently compatible with her not desiring to do what she believes that her fully rational self would desire that she do. As a consequence, it is difficult to see that it has been established that if an agent judges that an action is right for her to perform, then she is motivated to perform it on condition that she is rational, according to Smith's modified conception of rationality.

Second, an agent fulfilling the modified conception of rationality is compatible with her suffering from common forms of irrationality, such as weakness of will and depression. Assume that an agent desires to do what she believes that she, having a coherent set of desires, would desires that she, as she actually is, do. However, this is compatible with the agent, for example, lacking relevant true beliefs, being incapable of imagining available actions correctly, and having incoherent desires. The modified conception of rationality does thus not rule out depression and other instances of practical irrationality that constitute the target of Smith's original conception of this

34 In the argument, I consider (**I.3**). However, similar results emerge if we instead consider (**I.2**).

notion. As a consequence, the modified conception of rationality fails to explain that an agent who judges that an action is right for her to perform might not be motivated to perform it if she suffers from some common form of irrationality.

Finally, and more generally, the two above considerations underline that the modified conception of rationality referred to in (**I**) differs from the original conception of rationality as referred to in (**2**). As a consequence, it is difficult to see that (**2**) constitutes a moral analogue to (**I**) in the manner Smith's argument for (**2**) seeks to establish.

10. The Advice Model and the Objectivity of Moral Judgments

As a final point, I think it can be argued that the fact that Smith decides for (**P**.**2**) the Advice Model over (**P**.**1**) the Example Model might have significant implication for whether (**1**) the Objectivity of Moral Judgments has been established.

We have seen that Smith argues as follows in order to establish (**1**): (**R**) Rationalism and (**P**) the Platitude about Reasons, understood in accordance with the original conception of rationality, entail that whether it is right for an agent to perform an action is determined by considerations that are not relative to her actual desires but determined by objective circumstances. Thus, moral judgments are categorical which means that (**1**) has been established.

However, in case (**P**) is understood in line with (**P**.**2**) it can be doubted that Smith's analysis of moral judgments entails that moral judgments are categorical.[35] According to (**P**.**2**), what constitutes a normative reason for a less than fully rational agent depends on what her fully rational self would desire that she do, given how she actually is. As a result, it seems that there is no guarantee that all less than fully rational agents have the same normative reason if they find themselves in the same circumstances. This is so since what their fully rational selves would desire that their less than fully rational selves do depends on the desires of the latter. Notice that it may still be granted that all fully rational agents would end up with having converging desires so that they would all desire that a less than fully rational agent performs a particular action in a given situation. However, there is no guarantee that they would desire that two less than fully rational agents who find themselves in the same circumstances perform the same action if the desires of these agents differ in any significant respect. It follows that what normative reasons agents have might be relative to their actual desires. And from this it seems to follow that moral judgments are *not* categorical: different actions might be right for different agents to perform, although they find themselves in the same circumstances. The consequence seems to be that (**1**) the Objectivity of Moral Judgments cannot be sustained: moral judgments are not objective.

35 For related worries, see Hubin (1999), pp. 355–361 and Sobel (1999), pp. 137–147.

It should be stressed, however, that in Smith's view what would follow is not that moral judgments are not categorial. Rather, what would follow is that all moral judgments to the effect that actions are right are false. As noticed above, Smith maintains that if reasons were thought to be relative to actual desires, they would not be truly normative as they would be arbitrary. As a consequence, error-theory about normative judgments would be the result.

An obvious reply to this line of argument is to maintain that two less than fully rational agents have reasons to perform the same action in the same circumstances if 'circumstances' include actual desires. However, it is difficult to see that this understanding of 'circumstances' would make the account less conditional and save the categoricity of moral judgments in any substantive sense. Moreover, it seems a bit strained to maintain that an agent's desires are part of the circumstances of her situation.

It might further be replied that my argument merely shows that the demand that an agent's moral reasons should not be relative to her actual desires is too strong and has to be modified. Consider the squash example. It might be argued that what it is desirable that you do depends on your actual desire to smash your opponent in the face. To demand that what is desirable that you do should not be relative to your actual desire would be to demand that you should do something that is not desirable, namely to shake hands with your opponent with the result that you smash him in the face. Alternatively, it might be argued that to demand that what is desirable that you do should not be relative to your actual desire would be to demand that you should do something that you cannot do, namely to shake hands and leave the court without smashing your opponent in the face.

These arguments have already been considered above and I will not repeat my discussion of them. However, it should be recalled that (**P**.**2**) entails that in some cases where an agent's fully rational self forms a desire about what her less than fully rational self should do, she is sensitive to her less than fully rational self's desires even if the latter is able to perform an action that is considered desirable. To the extent that (**P**.**2**) has this implication, it seems to follow that what is right for an agent to do might be relative to her actual desires.

11. Conclusion

In his seminal book *The Moral Problem*, Smith sets out to solve the moral problem. In essence, the problem is that three intuitively plausible claims seem incompatible: (**1**) the Objectivity of Moral Judgments, (**2**) the Practicality of Moral Judgments, and (**3**) the Humean Theory of Motivation. Smith argues that these claims can be reconciled when properly understood. In this paper, I have primarily been concerned with Smith's efforts to establish (**1**) and (**2**) by considering his conception of practical rationality understood as coherence

between attitudes. Smith argues that (**1**) and (**2**) follow from (**R**) Rationalism and (**P**) the Platitude about Reasons. However, it was found that there are reasons to doubt this contention. In considering Smith's understanding of (**P**) it was argued that it entails implausible versions of (**I**) the Implication from Normative Beliefs to Rational Desiring, of which (**2**) is assumed to be a moral analogue. As a result, there are reasons to doubt that (**2**) has been shown to follow. Moreover, it was argued that Smith's understanding of (**P**) is incompatible with moral judgments being categorical. As a result, there are reasons to doubt that (**1**) has been shown to follow. Considering that there are reasons to question whether Smith's has established that (**1**) and (**2**) follow from his analysis of the relevant claims, there are also reasons to question that he has demonstrated that they are compatible. Thus, it can be doubted that the moral problem has been solved.[36]

References

Arruda, C.T. (2019). 'What the Humean Theory of Motivation Gets Wrong'. *Journal of Philosophical Research* 44: 157–178.

Björklund, F. et al. (2012). 'Recent Work on Motivational Internalism'. *Analysis* 72(1): 123–137.

Brink, D. (1997). 'Moral Motivation'. *Ethics* 108(1): 4–32.

Broome, J. (2013). *Rationality through Reasoning*. Oxford: Oxford University Press.

Copp, D. (1997). 'Belief, Reason, and Motivation: Michael Smith's The Moral Problem'. *Ethics* 108(1): 33–54.

Dancy, J. (1996). 'Real Values in a Humean Context'. *Ratio* 9(2): 171–183.

Francén, R. (2020), 'Reconsidering the Meta-ethical Implications of Motivational Internalism and Externalism'. *Theoria* 86(3): 359–388.

Gert, J. (2004). *Brute Rationality*. Cambridge: Cambridge University Press.

Gert, J. (2008). 'Michael Smith and the Rationality of Immoral Action'. *Journal of Ethics* 12(1): 1–23.

Hubin, D. (1999). 'Converging on Values'. *Analysis* 59(4): 355–361.

Johnson, R. N. (1997). 'Reasons and Advice for the Practically Rational'. *Philosophy and Phenomenological Research* 57(3): 619–625.

Kiesewetter, B. (2017). *The Normativity of Rationality*. Oxford: Oxford University Press.

Kolodny, N. (2005). 'Why be Rational?'. *Mind* 114(455): 509-563.

36 I am grateful to Joshua Gert for helpful comments on an earlier version of this paper.

Korsgaard, C. (1996). 'Skepticism about Practical Reason'. In *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, pp. 188–221.

Lillehammer, H. (1999). 'Analytical Dispositionalism and Practical Reason'. *Ethical Theory and Moral Practice* 2(2): 117–133.

Lord, E. (2018). *The Importance of Being Rational*. Oxford: Oxford University Press.

Markovits, J. (2014). *Moral Reasons*. Oxford: Oxford University Press.

Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.

Persson, I. (1995). 'Critical Notice of Michael Smith: The Moral Problem'. *Theoria* 61(2): 143– 158.

Sayre-McCord, G. (1997). 'The Metaethical Problem'. *Ethics* 108(1): 55–83.

Schroeder, M. (2007). *Slaves of the Passions*. Oxford: Oxford University Press.

Smith, M. (1989). 'Dispositional Theories of Value'. *Proceedings of the Aristotelian Society*, Supplementary Volume 63(1): 89–174.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Smith, M. (1995a). 'Internal Reasons'. *Philosophy and Phenomenological Research* 55(1): 109– 131.

Smith, M. (1995b). 'Reply to Ingmar Persson's Critical Notice of The Moral Problem'. *Theoria* 61(2): 159–181.

Smith, M. (1996a). 'Internalism's Wheel'. In *Truth in Ethics*, ed. Brad Hooker. Oxford: Blackwell, pp. 277–302.

Smith, M. (1996b). 'Normative Reasons and Full Rationality: Reply to Swanton'. *Analysis* 53(3): 160–168.

Smith, M. (1997). 'In Defense of The Moral Problem: A Reply to Brink, Copp and Sayre– McCord'. *Ethics* 108 (1): 84–119.

Smith, M. (1998). 'Ethics and the A Priori: A Modern Parable'. *Philosophical Studies* 92(1–2): 149–172.

Smith, M. (2009). 'Desires, Values, Reasons, and the Dualism of Practical Reason'. *Ratio* 22(1): 98–125.

Smith, M. (2012). 'Agent and Patients'. *Proceedings of the Aristotelian Society*, Supplementary Volume 112(3): 309–331.

Sobel, D. (1999). 'Do the Desires of Rational Agents Converge?' *Analysis* 59(3): 137–147.

Strandberg, C. (2007). 'Externalism and the Content of Moral Motivation'. *Philosophia* 35(2): 249–260.

Strandberg, C. (2011). 'The Pragmatics of Moral Motivation'. *The Journal of Ethics* 15(4): 341–369.

Strandberg, C. (2012). 'A Dual Aspect Account of Moral Language'. *Philosophy and Phenomenological Research* 84(1): pp. 87–122.

Strandberg, C. (2013). 'An Internalist Dilemma—and an Externalist Solution'. *The Journal of Moral Philosophy* 10(1): 25–51.

Strandberg, C. (2024). 'Neo-Humean Rationality and Two Types of Principles'. *Analytic Philosophy* 65(2): 256–273.

Swanton, C. (1996). 'Is the Moral Problem Solved?'. *Analysis* 53(3): 155–160.

Tresan, J. (2006). 'De Dicto Internalist Cognitivism'. *Noûs* 40(1): 143–165.

Williams, B. (1981). 'Internal and External Reasons'. In *Moral Luck*. Cambridge: Cambridge University Press, pp. 101–113.

John Eriksson
University of Gothenburg
john.eriksson@gu.se
Ragnar Francén
University of Gothenburg
ragnar.francen@filosofi.gu.se

Original Scientific Paper
UDC 17.022.1
161.2
Received: September 30, 2024
October 31, 2024
Accepted: November 01, 2024

PLATITUDES AND OPACITY: EXPLAINING PHILOSOPHICAL UNCERTAINTY

Abstract
In *The Moral Problem*, Smith defended an analysis of moral judgments based on a number of platitudes about morality. The platitudes are supposed to constitute conceptual constraints which an analysis of moral terms must capture "on pain of not being an analysis of moral terms at all". This paper discusses this philosophical methodology in light of the fact that the propositions identified as platitudes are not obvious truths – they are propositions we can be uncertain about. This, we argue, is a kind of *fundamental* philosophical uncertainty, and we develop an account of fundamental uncertainty (for both philosophical and other issues). The key to understand such uncertainty, on our view, is conceptual opacity – i.e., that the content and reference of concepts is not necessarily transparent to competent concept users. We argue that Smith's own view of conceptual analysis in TMP provides one plausible explanation of fundamental uncertainty. However, we also argue that another potential explanation is conceptual indeterminacy. If some fundamental philosophical uncertainties are best explained in this way, the implication is that there is no determinately correct analysis of the target terms and concepts.

Keywords: Platitudes · uncertainty · philosophical uncertainty · philosophical disagreement · conceptual opacity · conceptual indeterminacy

1. Introduction

One great merit, among many, of Michael Smith's *The Moral Problem* is its clear and eloquent presentation of the tension between the apparent practical and objective features of moral thought and talk. Consider the following three features:

1. Moral judgements of the form "It is right that I Φ" express a subject's beliefs about an objective matter of fact, a fact about what it is right for her to do.
2. If someone judges that it is right that she Φs then, ceteris paribus, she is motivated to Φ.
3. An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume's terms, distinct existences. (Smith 1994: 12)

 DOI: 10.5937/BPA2437081E

The first proposition concerns the objectivity of morality. The second feature concerns the practicality of morality. These features "pull in oppose direction from each other" (Smith 1994: 11) given the third proposition, i.e., the Humean theory of motivation. This is what Smith calls "the central organizing problem of contemporary metaethics" (Smith 1994: 11). No metaethical theory, at least as traditionally conceived, seems capable of making sense of the three propositions. Cognitivism makes sense of (1), but not (2). Non-cognitivism makes sense of (2), but not (1). Metaethics thus seems destined, in Mike Ridge's words, to try to "fit a square peg in a round hole" (Ridge 2014: 6).

Much of the last 30 years of metaethical theorizing has been driven by a desire to make sense of the apparent representational and non-representational aspects of moral thought and talk. Smith also claims that anyone who rejects one of the three propositions is "bound to end up denying something that seems more certain than the theories they themselves go on to offer" (Smith 1994: 13). Smith's aim in TMP is thus to outline a theory that captures all three propositions. Our focus in this paper is not to examine Smith's purported solution. Rather, our focus is more on its underlying methodology – another aspect of TMP that is masterfully clear.

Smith identifies what he calls "platitudes" about morality and builds his theory on them. "To say that we can analyze moral concepts, like the concept of being right, is to say that we can specify which property the property of being right is by reference to platitudes about rightness" (Smith 1994: 39). The three features that constitute the moral problem (see above) are supposed to be platitudes. Other examples of central platitudes mentioned include: "Judgements about rightness and wrongness are judgements about our reasons for and against acting." "If an agent believes that she has a normative reason to Φ, then she rationally should desire Φ." "When A says that Φ-ing is right, and B says that Φ-ing is not right, then at most one of A and B is correct." "Whether or not Φ-ing is right can be discovered by engaging in rational argument [...] and such arguments have a certain characteristic coherentist form." "Acts with the same ordinary everyday non-moral feature must have the same moral features as well." "What it is desirable that we do is what we would desire to do if we were fully rational; that what we have normative reason to do is what we would desire that we do if we were fully rational."

On Smith's view, "[a]n analysis of moral terms must in some way capture these various platitudes. It must do so on pain of not being an analysis of moral terms at all" (Smith 1994: 41). Many of the platitudes do indeed seem plausible, at least at first glance. But many of them (if not all) are also controversial. Furthermore, for many of the platitudes we, and probably many others, are uncertain whether they are correct. For example, we are drawn to motivational internalism, but are not fully sure whether it is in the end more plausible than externalism. Simply put, we are uncertain about a host of metaethical issues, including many of the propositions Smith calls

platitudes. But if the platitudes are what determines which metaethical theory is correct, what are we to make of this uncertainty? This is an instance of a more general issue. One central kind of philosophical task is to analyze and try to understand central philosophical phenomena and concepts: what is knowledge, truth, rationality, personal identity, free will, etc? If we are sometimes uncertain about the intuitions or propositions that form the foundation of those analyses, what causes such uncertainty and what are the implications (if any) of it. Perhaps it is a kind of conceptual uncertainty, but how should we understand such conceptual uncertainty?

The main aim of this paper is to advance an explanation of what it is to be fundamentally philosophically uncertain, in part by drawing on important remarks by Smith about the nature of conceptual analysis. This is done in the next section (section 2), where we first further specify the question to be pursued, and then provide our explanation. After that we turn to moral uncertainty: Smith (2002) argued that non-cognitivists cannot account for uncertainty in moral questions. We have previously suggested a non-cognitivist reply that parallels the ideas developed in section 2, and in section 3 we argue that replying to certain objections to this suggestion highlight the importance of Smith's (1994) views on conceptual analysis in an account of fundamental uncertainty. We end the paper (in section 4) by drawing out implications from our account of fundamental philosophical uncertainty for the views presented in TMP and for metaethics in general.

2. From conceptual opacity to fundamental philosophical uncertainty

2.1 Derived vs. fundamental uncertainty

It seems possible to be uncertain about virtually anything. A person can be uncertain whether a particular object is a chair or whether a particular philosophical thesis, e.g., motivational internalism, is correct. One question concerns what uncertainty *is*. We have nothing new to say about that. We may assume the common view that being uncertain whether p is to have a credence lower than 1 in p (i.e., believe p to a degree lower than 1). Instead, our question concerns how to understand a certain subclass of uncertainties: roughly, when you are uncertain whether x is F because you are uncertain what makes something an F. To single out the kind of uncertainties we are interested in, we introduce a distinction between fundamental and derived uncertainty. It can be illustrated with the trivial example of being uncertain whether an object is a chair.

Identifying an object, x, as a chair can be described as a two-part process. First, we observe (or in some other way come to think) that x has certain properties. Second, we are disposed to classify an object as a particular object,

e.g., a chair, by virtue of x displaying certain properties. Our classifications of an object as being, e.g., a chair, (or belonging to the same group of things more generally) is explained by us having and applying a classificatory disposition (or a "classificatory standard").

We can now say that a person's uncertainty regarding whether x is a chair is *derived* if it is caused by uncertainty whether x has properties that are chair-making according to their standard. Suppose that the person is blindfolded. This may make the person unsure whether the thing that they sit on has the kind of properties that make them disposed to classify it as a chair, e.g., whether it has a back.

A person's uncertainty regarding whether x is a chair is *fundamental* if it is not caused by uncertainty regarding X's properties or non-chair characteristics. Suppose that the person has full knowledge about all the properties of the object (except for being a chair of course), e.g., that it has four legs, a seat and a 10 cm high back. Despite knowing all the object's properties, the person may not be sure whether it is a chair. The person may not know whether to classify the object as a chair or a stool.

Although the example above may seem relatively uninteresting, it illustrates what seems to be a common phenomenon: sometimes we are simply unsure how a particular object is best classified. We may, for example, wonder whether the pope is a bachelor (even though we don't doubt he is an unmarried man), or whether an iPad (or any smartphone) is a computer. Our uncertainty about these matters is not caused by uncertainty about the properties of these things. Rather, despite being certain about these properties, we are uncertain whether x should be classified as a y. That we are sometimes uncertain in this way seems undeniable.

The distinction is relevant for uncertainty in philosophical matters as well. If we are uncertain whether someone's belief is a moral judgment (or perhaps rather a judgment of etiquette), that may be either because we are uncertain whether some property we take to be necessary to be a moral judgment is instantiated by the belief (derived uncertainty); or it may be because, even though we know every (other) property of this belief, we are uncertain whether that makes it a moral judgment (fundamental uncertainty). If we are uncertain whether S knows p, that may be because we are unsure whether S believes p (if we take that to be necessary to know p), or it may be because, even though we know everything about S's mental state, we are unsure whether it qualifies as knowledge. If we are uncertain whether S1 and S2 are numerically the same person, that may be because we are uncertain about the relation between S1 and S2, or because we are uncertain whether that relation makes them the same person.

To the extent that philosophers are concerned with understanding phenomena (and concepts) like moral judgments, knowledge and personal identity, with the aim of analyzing what it takes for them to be instantiated, the

fundamental versions of these uncertainties are central to philosophy. When we say that we are uncertain about some of Smith's platitudes, this seems to be fundamental uncertainty, for example. But fundamental uncertainty is also harder to understand than derived uncertainty. In the derived cases, it is pretty clear where the uncertainty comes from: uncertainty whether some specified condition for being a moral judgment, being knowledge, or being the same person, is fulfilled. But for fundamental uncertainty it is less clear. It is uncertainty about what it takes to be a moral judgment, be knowledge or be the same person – but where does it come from?

It should be noted that the difference between being derivatively vs. fundamentally uncertain whether x is a chair is not a difference between being uncertain *about different things*: both are to be uncertain whether x is a chair. It is holding the same belief – that x is a chair – to a degree lower than 1. Instead, the difference concerns the source. Our aim below is to describe different sources of fundamental uncertainty.

One key to understanding how fundamental uncertainty is possible is conceptual opacity. Suppose that S sees a chair-like object and that S knows about all of the object's properties, but still wonders if the object is a chair. If possessing the concept CHAIR (we will use small caps to refer to concepts) entailed full certainty about what is required of an object to be a chair (i.e., for an object to fall under the concept), then fundamental uncertainty would be impossible. However, fundamental uncertainty seems possible. Possessing a concept, C, does not necessarily entail that one is certain about what properties an object O must have in order for O to fall under C. We can therefore be uncertain whether something falls under C (or whether something is a C), even though we are certain about all other properties (i.e., apart from whether O is C) of O.

In the next three subsections, we will distinguish between three different ways in which fundamental uncertainty (both in general and in philosophy) can come about – partly depending on what explains conceptual opacity. In our discussion we will take our starting point in a metaethically more well-known issue: the open question argument. We'll argue that Smith's response to the open question argument provides the materials for one major source of fundamental uncertainty.

2.2 Fundamental philosophical uncertainty as metaphysical uncertainty

A well-known problem for naturalistic theories is G. E. Moore's open question argument. Consider, for example, the following analysis.

X is a grandmother = x is mother of a parent.

It seems self-contradictory to entertain the idea that x is a grandmother, but that x is not a mother of a parent. Similarly, "x is an mother of a parent, but is

x a grandmother?" seems to be a closed question. It seems to betray a kind of conceptual incompetence vis-à-vis the concept GRANDMOTHER. This suggests that the analysis above is right. Things seem very different if we consider moral concepts. Consider, for example, the following analysis.

X is right = x is approved of by the speaker.

By contrast to the previous example, it does not seem self-contradictory to entertain the idea that x is right, but that x is not approved of by the speaker. Similarly, "x is approved of by the speaker, but is it right?" seems to be an open question. Considerations along these lines led Moore to conclude that "right" "does not, by definition, mean anything that is natural; and it is therefore always an open question whether anything that is natural is [right]" (Moore 1903: 95/44).[1]

For a long time, philosophers were convinced that the open question argument showed that naturalism was a non-starter in metaethics. However, most philosophers now think that the argument relies on a mistaken view about conceptual analysis and the transparency of concepts.

A traditional way of thinking about the meaning of a word is in terms of a description that determines the referent and what a competent user of the word knows. For example, "grandmother" means "mother of a parent", where this provides necessary and sufficient conditions for the application of "grandmother". A competent user of "grandmother" knows this. Since this is part of what a competent user knows, this knowledge is also available a priori. This is why "x is a grandmother, but is x mother of a parent?" appears closed rather than open to a competent user of "grandmother."

Metaphysical naturalists understand the meaning of moral words differently. The meaning of "right" is given directly by its reference or what causally regulates our use of the word. The model for this idea is so-called natural kind terms, e.g., "water." "Water" refers to H_2O. However, that water is H_2O is not something that one can discover a priori, nor something one can know merely by virtue of being a competent speaker. Rather, it is something that we discovered a posteriori. The idea is that this is also the case for moral words. Hence, in order to find out the meaning of "right" we have to investigate what causally regulates our use of the word. Metaphysical naturalists think there is some natural property that causally regulates our use.

If this is how we think about the meaning of a word, it should not be very surprising that certain questions, e.g., "this is H_2O, but is it water?", appear open to us. The reference of a word or a concept is not necessarily transparent to competent users. People before the discovery were competent users of the word, but they did not know that water was H_2O. Similarly, it may be the case that "right" refers to a natural property, e.g., maximizing happiness, but this is not obvious to us nor something that we can find out by conceptual analysis.

1 Moore talked about goodness rather than rightness.

If this is so, it explains why, e.g., "x is N, but is x right?", appears to be an open question even for competent users (even if rightness actually is N).

This also provides a simple explanation of fundamental uncertainty. The conclusion that it can remain an open question to competent users of "water" and "right" whether water is H_20 and right actions are N, is, in effect, the conclusion that such competent speakers can be uncertain of the correctness of these claims. If a causal theory of reference is correct about these concepts, this is also an instance of fundamental rather than derived uncertainty (according to the distinction above). Since there is no specific description that competent users need to associate with water, a competent user can know all the properties of a substance but still be uncertain whether it is water. Hence, when a person is uncertain in this way about whether x is water, the uncertainty is not necessarily derived from uncertainty about whether x has some specific property that is "water-making" according to the description or standard that she associates with water.[2]

The uncertainty in question would be a form of metaphysical uncertainty, rather than conceptual uncertainty. What one would be uncertain about is, for example, the metaphysical nature of water. To become less uncertain, this is what one would have to investigate.

Given certain philosophical views, some fundamental philosophical uncertainties may be like this. Suppose, for example, that we think that mental state kinds – e.g., beliefs and desires – are natural kinds (or that our concepts BELIEF and DESIRE function sufficiently much like natural kind concepts do according to the theories described above) then if S is uncertain whether beliefs are necessarily evidence-sensitive, this can be understood as a metaphysical uncertainty. Likewise, if we think that MORAL JUDGMENT is a natural kind concept, when we are uncertain whether moral judgments necessarily come with motivation (in practically rational people), this can be seen as a metaphysical uncertainty.

This is not how Smith (1994) would think about uncertainty regarding the platitudes, however. For he sees them as *conceptual* constraints on the correct analysis. Given this, such uncertainty must be understood differently. Here too, we may start with the open question argument and Smith's reply to it.

---

2 However, if neo-descriptivists like Jackson (1998) are correct, then we should rather think of it as a kind of derived uncertainty. Even if there is no molecular structure-description that a competent speaker/concept user needs to associate with water, there are other kinds of descriptions. For example, that water is the clear, potable liquid found in lakes etc, for short, the "watery stuff" we are actually acquainted with. Hence, if S is competent user of "water" and is uncertain whether a certain substance is water even though S knows all about that substance's chemical composition, that will be because S is uncertain whether that substance has one property that she takes to be water-making, namely if it has the property of being the same kind of substance as the watery stuff found in lakes etc. This makes it derived uncertainty.

2.3 Fundamental philosophical uncertainty as conceptual uncertainty

By contrast to metaphysical naturalists, Smith defends definitional naturalism, i.e., "the view that we can define moral terms exclusively in terms apt for describing the subject matter of the natural and social sciences" (Smith 1994: 35). This view is more in line with the traditional view of meaning. Indeed, one may hypothesize that this is the kind of view that Moore had in mind. Smith has a simple and ingenious argument purporting to show that the open question argument is not a problem for this kind of naturalism either. The argument turns on the nature of conceptual analysis.

> ...consider the enterprise of giving analyses quite generally. It is a familiar fact about analyses that a concept C* may constitute a correct analysis of a concept C despite the fact that it is possible to think that x falls under C* and yet also, apparently coherently, entertain the possibility that x does not fall under concept C. (Smith 1994: 36)

Smith considers examples like "red," "knowledge," and "intentional action." "Red," for example, might be analyzed as "having the property that causes objects to look red to normal perceivers under standard conditions." However, this analysis does not have to be obvious even to competent users of "red."

> The point I am making here does not require the assumption that we can correctly analyze the first in terms of the second in each case. The point is rather that the mere fact that it is an open question whether we can – something that can hardly be denied given the number of pages devoted to discussing these suggestions in the philosophical literature – is not already enough to show that we cannot. (Smith 1994: 36)

This idea also constitutes a response to what is known as the paradox of analysis.

> Paradox of analysis – when we are looking for an analysis of a concept C, we are looking for a concept C* that will tell us something new and interesting about C, something we don't already know. The claim that C is analytically equivalent to C* must therefore be unobvious and informative in some way. But C* must also really be analytically equivalent to C (Smith 1994: 37).

Smith convincingly argues that the paradox of analysis is "an artifact of bad views about the nature of conceptual analysis" (Smith 1994: 37). We simply don't have direct access to the structure of our concepts. This is why an analysis is not transparent. Consider Smith's view about concept acquisition.

> ...in acquiring a concept C we come to acquire a whole set of inferential and judgemental dispositions connecting facts expressed in terms of the concept C with facts of other kinds. A statement of all of these

> various dispositions constitutes a set of platitudes surrounding C. And an analysis of a concept is then best thought of as an attempt to articulate all and only these platitudes. (Smith 1994: 37–38)

On Smith's view, the meaning of "red" and "right" is to be understood in terms of the kind of descriptions that determine the reference of the words and that competent users know. However, there is an important ambiguity in the use of "know" here. It is one thing to know *how* to use a word. This is the kind of knowledge that competent concept users have: they are disposed to use the concept in the right way. It is quite another thing to know *that* a particular description takes us to the referent. This requires obtaining a correct description of the dispositions of competent users. For a competent user, coming to know the correct description – and thereby the analysis of the meaning – requires investigating one's own patterns of dispositions of use in a wide range of relevant actual and hypothetical cases. Thus, before such an investigation, the description (or analysis) is not transparent even to competent users. This is why asking, e.g., "x is N, but is x right?" appears to be an open question.

We find this view very appealing. Since it implies that conceptual content is opaque even to competent users, it also provides an explanation of fundamental uncertainty, distinct from the one made available by causal theories of reference. Consider, a person S who considers whether an object must have legs to be a chair. S may not be sure what to think. The explanation may be that S has not gone through, and realizes that she has not gone through, the process of finding out whether there are actual or hypothetical examples of chair-like objects without legs that S is disposed to classify as chairs.

This is a kind of fundamental uncertainty. It is not caused by thinking that having certain properties is necessary to be a chair, and being uncertain whether those properties are instantiated. Rather, it is caused by uncertainty about which properties are necessary to be a chair. Which is made possible by conceptual opacity. In contrast to the kind of fundamental uncertainty discussed in the previous subsection, this is not metaphysical uncertainty but conceptual uncertainty. It does not stem from uncertainty of the metaphysical nature of the things referred to by the concept CHAIR, but rather from uncertainty about what the conceptual constraints of that concept are.

It seems plausible that some philosophical uncertainty is of this sort. One may be uncertain whether motivational upshot is necessary (given certain conditions) for having a moral judgment, or whether psychological continuity is necessary for personal identity, simply because one realizes that there might be cases that one has not yet thought of where one would be disposed to judge in conflict with those propositions. Indeed, Smith's own view would imply that we can be uncertain in this way about the platitudes that form the ground of his theory in TMP. The platitudes about a concept (e.g., rightness), on Smith's view, are precisely statements of the inferential

and judgmental dispositions involved in mastery of the concept. Since one can master a concept without knowing how to spell out these dispositions, one can master a concept while still being uncertain about whether the (actually correct) platitudes are correct.

There is also a related reason why conceptual content is opaque. Our responses to cases are sometimes influenced not only by our conceptual competence, but also by distorting factors, such as our not grasping all relevant details of the scenario or wishful thinking. This means that even if a competent user has considered all relevant cases in relation to some concept (and her intuitive responses to these), this does not guarantee that she knows the correct analysis of the concept since her response may be due to distorting factors. This is another source of fundamental uncertainty.

It should be noted that we are not claiming that not having considered all relevant cases (and not having checked for distorting factors) automatically gives rise to this sort of uncertainty. Or that a person's uncertainty is always proportional to the degree to which she has checked these parameters. Let us take an example. Before Gettier presented counterexamples to the tripartite analysis of knowledge, some philosophers may have been completely certain that justified, true belief were sufficient for knowledge. This obviously was not because they had considered every relevant case. Other philosophers may have been less than certain even before Gettier's paper. To what extent a person thinks that there may be relevant cases she has not thought of, or that she might have misconstrued certain cases, will probably be due to things like tendencies to dogmatism versus openness to being wrong.

The example of Gettier cases also illustrates another thing. Considering new cases can not only relieve uncertainty but also induce it. If one is first pretty sure that knowledge is justified true belief, then encountering Gettier cases may make one certain that this is not the case. But alternatively, it may make one become uncertain whether it is the correct analysis, for one may suspect that there could be some way of understanding the examples that is consistent with the tripartite analysis. Again, this depends on conceptual content being opaque. Given Smith's view of conceptual analysis, we also have a way to remove uncertainty, of course. We simply have to consider more sophisticated scenarios (and make sure to avoid distortions) that help us explore the inferential and judgmental dispositions.

2.4 Fundamental philosophical uncertainty as conceptual indeterminacy

If all conceptual uncertainty is of the kind discussed in the previous subsection, then ideally, if we could consider all relevant cases (and get rid of all distortions), we would also get rid of all conceptual uncertainty. We think this picture ignores one important source of fundamental conceptual uncertainty, however. It presupposes that our concepts are determinate in

the sense they will give rise to determinate responses in every hypothetical scenario that we can consider. We are not sure that this is correct. When we investigate our inferential and judgmental disposition by considering real and hypothetical cases, it may turn out that we don't know what to think, i.e., we are not (clearly) disposed to judge whether a particular object, e.g., a floating seat, is a chair or not. When we consider a case like this, our concept (or our classificatory standard) does not apply. The explanation, we think, is that our concepts (or our classificatory standards) are indeterminate regarding some cases. Instead of having a clear intuition regarding whether a floating seat is a chair we are disposed to vacillate (is it a chair or not?) or remain more radically undecided (we simply don't know how to classify the object).

Of course, sometimes when someone is undecided whether to call a certain object a "chair" or not, the explanation is that she does not fully master the CHAIR concept and the meaning of the term "chair". But what we are suggesting here is that there are also cases when the concept itself is indeterminate, so that even people that master the concept are disposed to remain undecided. In fact, we think that there are cases like this for most, if not all, concepts. Consider the following passage from Peter van Inwagen.

> To specify the meaning of a predicate is to give a set of instructions for its application, and it is well-nigh impossible for a set of instructions to cover every possible situation; in consequence, no matter how carefully we specify the rules for using some new predicate that we propose to introduce into our language, there will almost certainly be possible cases in which it is indeterminate whether that predicate applies. (And, as many writers have pointed out, when one introduces a new predicate, there will normally be good, practical reasons for leaving it indeterminate whether it applies in possible cases in which one could render its application determinate. As Lewis has said, no one has ever been fool enough to try to specify the precise portion of the surface of the earth as the referent of 'the outback'. It would seem, therefore, that all or almost all predicates will admit of possible borderline cases; (van Inwagen 2009: 4)[3]

If we connect this to Smith's idea of concepts and concept acquisition, the idea is that for many concepts, the inferential and judgmental dispositions that we acquire when we acquire a concept don't cover every possible situation. For most concepts, there will be cases where there has been no need for

3 Other philosophers have also argued for conceptual indeterminacy, let us mention two examples. Waismann (1945) argued that many concepts are open textured, which roughly means that there are (and always will be) genuine borderline cases regarding their application, that is, possible cases where there is no unique correct answer to whether the term applies or not. More recently, but in a similar vein, Ludlow (2014) has argued that for many words in natural languages it holds that "even after a millennium of shared usage the meaning is quite open-ended" (p. 1) and that "word meanings themselves are dynamic and massively underdetermined" (p.3).

our linguistic/conceptual community to decide on whether they do or don't fall under the concept. For example, when "chair" was introduced into the English language, it was not meant to cover examples involving floating seats, and as the meaning of that term has evolved, such cases have not been common or salient enough to force a sharpening of the meaning to make the term (and corresponding concept) to either apply or not apply. It is therefore indeterminate whether "chair" (and the corresponding concept chair) applies to floating seats.

A plausible case can be made that the same thing applies to many philosophically relevant terms. For example, it may be argued that our concept PERSON did not develop in contexts where there was a pressure to determine whether personal identity survives in branching cases. As a consequence, our everyday concept PERSON is indeterminate regarding whether personal identity survives branching. We see no reason not to think that this is true for many (if not most) philosophical concepts. This gives us a third source of fundamental uncertainty, both generally and in philosophy.

It should be noted that this source of fundamental uncertainty also presupposes a kind of conceptual opacity. Suppose that the CHAIR concept is indeterminate with regard to floating seats, and that conceptually competent people are therefore undecided or vacillate when they consider whether such objects are chairs. Why would this undecidedness cause uncertainty about whether these objects are chairs, rather than causing certainty that they neither are chairs nor not chairs (i.e., that the concept is indeterminate)? It is experienced as uncertainty because it is not transparent that the undecidedness or vacillation comes from conceptual indeterminacy, the CHAIR concept does not wear its indeterminacy on its sleeve. Rather, people tend to presuppose that most objects either are chairs or not. This means, that although what we have is indeterminacy at the level of concepts (or classifications), this is not how people experience it. Instead, they experience it as uncertainty at the object level – is it a chair or not? (We will return to the issue of indeterminacy and opacity in a response to an objection in the next section.)

We are not claiming, however, that indeterminacy always causes uncertainty. Sometimes it is experienced precisely as indeterminacy (instead of uncertainty). One may, for example, think that there is no correct answer whether an iPad is, or is not, a computer, that it is just a case of conceptual unclarity. What determines when indeterminacy is experienced in one of these ways rather than the other probably depends partly on what kind of concept we are dealing with (we may e.g., be less prone to think of natural kind concepts as indeterminate than artifactual kind concepts), but may also vary from person to person.

We should also distinguish between indeterminacy-caused uncertainty and a related phenomenon. When there has been no need for our community to fix a concept's and term's extension regarding a certain class of cases so

that it determinately applies or not, one of two things (or both) may happen. The first is what we have described above: competent concept users become disposed to be undecided in their application. But it has also been argued that this leaves it open for competent users to be decidedly disposed to judge differently: e.g., some may judge that the floating seat is a chair, others may judge that it isn't, without either judgment indicating conceptual incompetence (Francén 2022). If this is the case, the lack of a joint determinate way of classifying these cases does not entail that all competent users lack a disposition to classify in a determinate way. It is only when individual competent users lack such a disposition that it can cause uncertainty.

We have now described three distinct causes of fundamental uncertainty, both in general and in philosophy. In section 4 below, we will discuss the metaethical consequences of the different variants of fundamental philosophical uncertainty described above. Before that, however, we will discuss how we have previously used the kind of account described above to reply to an objection advanced by Smith (2002) to the effect that non-cognitivists cannot account for moral disagreement.

3. Non-cognitivism and moral uncertainty

A few years after TMP, Smith argued that non-cognitivists cannot account for degrees of certitude in moral judgments (Smith 2002). If P judges that torturing an innocent person is always wrong, P may be more or less certain about this, but P can be more certain about this than that lying is always wrong. The problem is that if moral judgments are desires, the *strength* of these desires plausibly corresponds to *importance* – e.g., to thinking that an action is more or less immoral. Then there is no gradable feature of desires left that can constitute degrees of *certitude*.

In our (2016) we presented a solution on behalf of non-cognitivists which is in line with the general view about fundamental uncertainty presented above – what we called "the Classificatory Account". Our aim here is not to argue that this is the best account for non-cognitivists, but rather to show that what we have said above (e.g., about Smithian opacity of conceptual content), provides replies to objections that have been raised against it. Those replies also shed further light on the general view presented above.

According to the Classificatory Account, non-cognitivists should say that (ordinary non-normative) beliefs necessarily accompany moral judgments[4], and moral uncertainty is located in these beliefs.[5] Bykvist & Olson (2009) had previously argued that such views fail to account for *fundamental* moral uncertainty, and the Classificatory Account was proposed as a solution to this problem.

4 Or form part of them, as a form of hybrid expressivism, but we ignore that option here.

5 For an earlier solution of this sort, see Lenman (2003).

What does the Classificatory Account say? Suppose that non-cognitivists have an independent account of which non-cognitive attitudes that (at least partly) constitute moral judgments – i.e., the relevant sorts of approvals and disapprovals. Different people respond with moral (dis)approval to different features of acts, e.g., the consequences for well-being, the intention behind the act, etcetera. This can be described as follows: Each moral judge has some classificatory standard such that (i) it classifies acts that have certain (non-moral, descriptive) properties together, and (ii) those classifications regulate her formation of moral approvals. The classificatory standard that has this function in a person's psychology (the function of regulating her moral approvals) we can call her *moral rightness-standard*.

On this view, although moral judgments are constituted by noncognitive attitudes, beliefs are always involved in their formation. As noted, when a person P morally approves of an act A, she always does that based on some descriptive property F (most often a complex combination of other descriptive properties) of A. This means that in the process of forming her approval, P forms the belief that A has property F. Which complex property F is might not be known by P – to know this we need to analyze her dispositional pattern of morally approving of acts based on their features, i.e., her moral standard. This is analogous to Smith's view that mastering a concept (having the dispositions) does not entail knowing how to describe it.

Let us call the belief that figures in a person's formation of moral rightness-judgments her "moral approval-regulating belief". These beliefs are not themselves *moral beliefs*, but beliefs with a descriptive content that regulate the formation of the approvals that constitute moral rightness-judgments. The content of the moral approval-regulating beliefs (that is: which property F is) will vary from person to person, depending on their moral standards. Let us take a simple example. Suppose that P has a utilitarian moral standard. She might not know this herself, but if we were to analyze her dispositional pattern, it would turn out that she is disposed to morally approve of an act if, and only if, she thinks that the act maximizes wellbeing. For her, then, it is the belief that an act maximizes wellbeing that is moral approval-regulating. She might not be aware of having a belief with this content (or any belief about the act's descriptive features) when she forms her moral judgment. But anyway, the belief is there.

According to the Classificatory Account, P's being more or less certain that A is right, is for P to hold the moral approval-regulating belief that accompanies her moral approval – i.e., the belief that A is F – with more or less certitude. Derived and fundamental uncertainty are distinguished in the same manner as for non-moral uncertainty. The uncertainty is derived when it is caused by (i.e., derived from) uncertainty about whether A has some of the features that are right-making according to A's moral rightness-standard: e.g., does A cause suffering? It is fundamental moral uncertainty when it is not derived from such uncertainty.

Here, just like in the general view presented above, we distinguished between three ways in which such fundamental uncertainty can arise. First, P may be uncertain whether all white lies are morally wrong because her wrongness-standard (the classificatory standard that drives her disapproval formation) is indeterminate here: the standard is such that she is not clearly disposed to judge them wrong nor not wrong, or she oscillates between the two. Second, since she may not have explicit knowledge of whether she is disposed to classify (and thereby (dis)approve) of) all instances of an act type, e.g., the killings of innocent, she may not be entirely sure that all such acts are wrong. Third (and corresponding to metaphysical uncertainty above), due to projectivist inclinations (often adduced by non-cognitivists to explain the experience of moral right and wrong being in the world), she may have a sense that, even though her own rightness-standard might be determinate, and she is not uncertain about it, it might not align with which actions are actually right. These three factors can give rise to the moral approval-regulating belief being held with less than full certitude, without that incertitude being derived from incertitude about whether any particular rightness-making (descriptive) feature is in place. Consequently, this is an example of fundamental moral certitude.

We will now consider two objections to this view of fundamental moral uncertainty, and argue that they fail once the opacity of concepts is acknowledged.

3.2 Objections and replies

First objection. As explained above, our suggestion was that non-cognitivists should say that degree of moral uncertainty is located in a belief that necessarily accompanies the (dis-)approval. For example, for a person with a utilitarian moral standard, the relevant belief concerns whether the action in question maximizes wellbeing. This idea is the target of one of Bykvist & Olson's (2017) main objections:

> But this will not give us fundamental moral uncertainty, since uncertainty about whether an act maximizes happiness is, according to Eriksson and Francén Olinder's own account, not a case of fundamental moral uncertainty. As they put it, in the abstract of their paper, '[a] person is *derivatively* [that is, not fundamentally] uncertain about whether an act is, say, morally wrong, when her certainty is at bottom due to uncertainty about whether the act has certain non-moral, descriptive, properties, which she takes to be wrong-making.' (Bykvist and Olson 2017: 796)

Based on this, Bykvist & Olson dismiss this view about the location of moral uncertainty. Taken by itself, this objection simply seems to miss that on the proposed view, when someone is uncertain in her belief that p, it is the *source*

of the uncertainty (i.e., what it "is at bottom due to"), and not the content of the belief, that determines whether the uncertainty is fundamental or derived. P's uncertainty that x is a chair can be derived *or* fundamental. The content of the belief is the same in both cases. This is also the case for moral uncertainty.

But Bykvist & Olson probably have a related complaint in mind: that the proposed view doesn't escape their earlier objection towards Lenman's (2003) view, which also locates moral uncertainty to a belief with a descriptive content. According to their objection, Lenman's view fails to account for *moral* uncertainty (rather than some descriptive uncertainty) since one can hold any descriptive belief about action A with full certainty, but still be morally uncertain about e.g., whether A is wrong (Bykvist & Olson, 2009). Directed to the Classificatory Account, the objection is as follows. On that account, if P has a utilitarian standard, the content of her moral approval-regulating beliefs concerns whether an act maximizes well-being. When P is uncertain whether A is right, this means that she holds her moral approval-regulating belief – with the content that A maximizes wellbeing – to a low degree. Bykvist & Olson's objection, then, is that this precludes that P can be fully certain that A maximizes wellbeing but still uncertain whether A is right.

But given Smith's ideas about the opacity of the content of concepts, this objection is misguided. If the correct analysis of the contents of our concepts and beliefs are not always transparent to us, we can be both ignorant and mistaken about them. Thus, P can be unaware that her moral approval-regulating belief has the content that A maximizes wellbeing (since her dispositional patterns may be opaque to her). Hence, she can be fully certain that A maximizes wellbeing, but still uncertain in her moral approval-regulating belief. Bykvist and Olson's objection therefore fails.

We can illustrate this point by returning to a case of non-moral uncertainty. Suppose P believes that x is a chair but is not fully certain. That uncertainty consists in P holding the belief *that x is a chair* to some degree lower than 1. Plausibly, this belief is where the uncertainty is *always* located – irrespectively of whether the uncertainty is derived or fundamental. If P's uncertainty springs from uncertainty about some of x's characteristics (e.g., does it have a back? can you sit on it?), then it is derived. If P is *certain* about x's non-chair characteristics but still uncertain whether a thing like that (with those characteristics) count as a chair – then her uncertainty is fundamental chair uncertainty. Hence, it seems quite possible that one can be fully certain about all non-chair characteristics of an object, but still be uncertain whether it is a chair. However, this may seem to have paradoxical consequences, analogous to the problems that Bykvist & Olson point out for moral uncertainty, given that the chair-concept can plausibly be analyzed in terms of non-chair properties. Suppose the chair-concept is such that being a chair is to have properties F_1-F_n. Consequently, believing that, and being

(un)certain whether, x is a chair is in effect to believe that, and be uncertain whether, x has F_1-F_n. But then it would seem to follow that you cannot be fully certain that x has F_1-F_n and still be uncertain whether x is a chair – for the latter uncertainty would be uncertainty whether x has F_1-F_n.

But the paradoxical impression is dissolved once we recognize that the correct analysis of the contents of our concepts and beliefs are not always transparent to us. If I don't know that CHAIR is the concept of having F_1-F_n, then I may be (i) uncertain whether x is a chair, while (ii) believing with certainty that x has F_1-F_n – even though (i) is, in effect, (though I'm not aware of this) uncertainty in a belief with the content *that x has* F_1-F_n. Simply put: due to the opacity of content, even though *x is a chair* and *x has* F_1-F_n has the same content, you can believe one but disbelieve (or be uncertain about) the other. Likewise, due to conceptual opacity, even if Ps moral approval-regulating belief has the same content as the belief that x maximizes happiness, one can hold the latter but not the former with full certainty. This illustrates why Bykvist & Olson's objection to the Classificatory Account fails: it ignores the kind of conceptual opacity that grounds Smith's reply to the open question argument in TMP.

Second objection. Ridge (2018) argues that indeterminacy cannot be part of an explanation of uncertainty. The ambivalence we have when applying an indeterminate or vague concept to a borderline case, is plausibly distinct from uncertainty. In contrast to cases of uncertainty, when we think of a borderline case of, say, baldness, we don't take ourselves to be ignorant of the real state of affairs and don't have a "sense that further investigation might in principle help firm up our view" (Ridge 2018: 3329).

Ridge's point is relevant, but just like the previous objection it ignores the opacity of conceptual matters. When we are completely aware that our ambivalence about a case is due to conceptual indeterminacy (as sometimes is the case in clear cases of vagueness), we will indeed not think that there is a fact of the matter to be ignorant of – conceptual indeterminacy implies, after all, that it is indeterminate whether the concept applies. However, indeterminacy can be a source of uncertainty because we are often not aware that our ambivalence is due to indeterminacy. When a concept is not obviously vague (and sometimes even then) we expect there to be determinate answers to each question of whether x falls under the concept or not, and then we often (mistakenly) interpret our own lack of determinate intuitions about the classification of some case as ignorance about the (determinately) correct classification, also when it is due to the concept being indeterminate. For example, we expect each object to either be a chair or not. Hence, our ambivalence about how to classify an object turns into uncertainty about whether it is a chair. (This is what we argued in section 2.4 above.) Similarly, we expect acts to be either morally wrong or not.

To sum up, through our responses to the objections above, we have argued that failure to acknowledge conceptual opacity will make certain kinds of fundamental uncertainty seem impossible. This highlights the importance of recognizing conceptual opacity to fully understand the phenomenon of fundamental uncertainty. As modern discussions of among other things the open question argument have shown, just because something is a conceptual truth, it is not an obvious truth even to competent users of the concept in question. Smith's ideas in TMP make this eminently clear, but these ideas also help us respond to objections regarding fundamental uncertainty for non-cognitivists – a problem that originates from Smith.

4. Implication for metaethics and the moral problem

The content of concepts, even if a priori, is not transparent – not even to competent users. Smith used this to explain why certain moral questions seem open, even if "right" and "wrong" can be analyzed in wholly naturalistic terms. We have argued that this also explains (at least some cases of) fundamental philosophical uncertainty. In this final section, we draw out some implications of this idea for metaethics, The Moral Problem and philosophy more generally.

The underlying methodology of TMP is to work from platitudes about, e.g., rightness, to an analysis. Some purported platitudes support the objectivity of morality. For example, "When A says that Φ-ing is right, and B says that Φ-ing is not right, then at most one of A and B is correct." This proposition may seem plausible, but if the contents of concepts are opaque in the way Smith argues, there is room for uncertainty. On the one hand, this is a good thing for Smith: he can hold that they are conceptual constraints, even if some people are not certain that they are correct or even reject them.

On the other hand, it complicates the methodological picture. For given the opacity, it may turn out that even if e.g., the proposition above at first appears (to Smith and to many of others) to be an obvious platitude, it might not really be a platitude (i.e., a conceptual constraint). For there may be cases about which competent concept users are disposed to make judgments that contradict the proposition. Indeed, regarding this particular platitude, uncertainty rather than certainty on Smith's behalf seems to be warranted. Sarkissian et al. (2011) challenged the claim that people have objectivist intuitions in the moral domain by presenting people with hypothetical cases. Their question concerned precisely whether people judge that, in a situation where A judges that an action is immoral and B judges that it is not immoral, one must be wrong. They found that "people's intuitions take a striking relativist turn when they are encouraged to consider individuals from radically different cultures or ways of life" (Sarkissian et al 2011: 500). The general point here is the following: given that we accept conceptual opacity of

the sort Smith argues for, and given philosophers' track record of coming up with counterexamples to purported platitudes, *uncertainty* about the status of seeming platitudes seem warranted.

This might seem trivial: of course philosophers should be open to counterarguments – we should not be dogmatists. But we think the implications are worth pondering. For Smith, to seriously keep open the possibility that a purported platitude is, in fact, not a platitude would be to seriously hold open that a radically different analysis is correct (perhaps a non-realist one). To be clear, this is not to say that there is no philosophical value to the method of suggesting that certain propositions are platitudes. Quite the opposite. We must start with propositions that we find plausible, but we may, through philosophical reflection, realize that they were mistaken. The clear way that Smith presents the (purported) platitudes is an example of this. It has pushed the debate regarding many different issues forwards in ways that few other books have done.

Conceptual opacity is also connected to disagreement. At the outset of TMP Smith suggests a neat explanation of the wide-ranging disagreement in metaethics, namely that it is an effect that the practicality and objectivity features of morality seem to pull in different directions given the Humean theory of motivation (Smith 1994: 4–5). However, this only accounts for some of the disagreement. For every platitude that Smith considers, there will be philosophers who disagree with it. Just think about the number of papers devoted to the objectivity and practicality of morality. Not only do intuitions about platitudes differ among philosophers, they also have different intuitions in response to the scenarios that, e.g., Sarkissian et al. presented, and scenarios about moral judgments that come apart from motivation. That is, the kind of scenarios which we should consider to tease out which inferential and judgmental dispositions competent speakers have. One explanation of this refers to the kind of conceptual opacity Smith highlights, i.e., the difficulty of investigating our interferential and judgmental dispositions. If this is the only explanation, then a process involving a more thorough investigation of more and more scenarios should move us toward certainty and agreement.

But we don't think that such a process will eliminate all disagreement, since we see no reason to think that for most concepts, conceptual competence involves being disposed to determinately and unambiguously judge all scenarios. That is, we find it likely that many concepts central to philosophical theorizing are indeterminate. When a philosophical concept is introduced, paraphrasing van Inwagen's point, it seems implausible to think that it comes with a set of instructions that cover every possible situation. When we encounter certain hypothetical cases, it will therefore be indeterminate whether the concept applies. However, as we suggested above, this is often opaque. When we lack a determinate disposition, we will therefore experience uncertainty about the true extension of the concept in the case at hand. One reaction to such uncertainty is to sharpen one's standard so that the

concept applies to the case in a determinate way. This way one rids oneself of uncertainty about the case at hand (and maybe about the correct analysis of the concept, at least until one comes across another hypothetical case where the concept does not apply).[6] Given that the opacity of the concept hides the indeterminacy to start with, this will not be experienced as sharpening the concept. Rather, it will be experienced as a discovery or as finding something out. However, what really happens is that our concept undergoes a change.

We suspect that this is often what happens in philosophy. Part of the reason why we have different intuitions is that we endorse different theories. When we encounter a scenario where our concepts don't apply, we sharpen them to fit with our larger theoretical frameworks. (And when we don't have a theoretical framework to uphold, on pain of not being able to take a stand in the debate at hand, we need to set down our foot.) This results in philosophers coming to have slightly different concepts. This, in turn, provides at least a partial explanation of why there is so much disagreement and why these disagreements are (next to) impossible to resolve.[7]

Of course, we have not *shown* here that philosophically or metaethically relevant concepts are indeterminate in such a way that the answer to some philosophical questions is indeterminate. Some philosophers have argued this, however. In metaethics Gill (2009) is the clearest example. More generally, Unger (1984) argued that something like this holds for many philosophical questions, e.g., questions about knowledge, free will, and causation. These may be radical theses, but we think they should be taken seriously, especially given some of the assumptions Smith starts from. Given that we think of conceptual contents in terms of dispositions of competent users, then in light of the kinds of concerns raised by van Inwagen and others, the claim that competent users will have developed determinate dispositions for all kinds of cases is something that needs to be argued for, rather than assumed.

References

Bykvist, K., & Olson, J. (2009). Expressivism and Moral Certitude. *The Philosophical Quarterly* 59/235: 202–15.

Bykvist, K., & Olson, J. (2017). Non-Cognitivism and Fundamental Moral Certitude: Reply to Eriksson and Francén Olinder, *Australasian Journal of Philosophy*, 95/4: 794–799

6 On Waismann's (1945) view, the open texture of concepts cannot be removed: if we sharpen the concept with regard to one kind of case, there will always be other "unfamiliar cases" where the concept is indeterminate.

7 As noticed above, it has also been argued that even disregarding such philosophy internal pressures, people may have developed concepts that are sharpened in different ways regarding the scenarios that the "joint instructions" don't cover (Francén 2022).

Eriksson, J., & Francén, R. (2016). Non-Cognitivism and the Classification Account of Moral Uncertainty, *Australasian Journal of Philosophy* 94(4), 719–35.

Francén, R. (2022). Mananas, Flusses and Jartles: Belief Ascriptions in Light of Peripheral Concept Variation. *Philosophical Studies* 179, 3635–3651. https://doi.org/10.1007/s11098-022-01859-6

Gill, M. B. (2009). Indeterminacy and Variability in Metaethics. *Philosophical Studies* 145, 215–234.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford, Oxford UP.

Lenman, J. (2003). Non-Cognitivism and the Dimension of Evaluative Judgment. Brown Electronic Article Review Service, (ed.) J. Dreier and D. Estlund, URL= http://www.brown.edu/Departments/Philosophy/bears/0301lenm.html

Moore, G. E. (1903). *Principia Ethica.* Cambridge, Cambridge University Press

Ridge, M. (2014). *Impassioned Belief*. New York, Oxford University Press.

Ridge, M. (2020), Normative Certitude for Expressivists, *Synthese* 197, 3325–3347 https://doi.org/10.1007/s11229-018-1884-7

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Smith, M. (2002). Evaluation, Uncertainty, and Motivation. *Ethical Theory and Moral Practice* 5(3), 305–20.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk Moral Relativism. *Mind and Language* 26, 482–505.

Unger, P. K. (1984). *Philosophical Relativity*. Minneapolis: University of Minnesota Press.

Van Inwagen, P. (2009). Indeterminacy and Vagueness: Logic and Metaphysics. *European Journal for Philosophy of Religion* 1(2), 1–19.

Waismann, F. (1945). Symposium: Verifiability. *Proceedings of the Aristotelian Society, Supplementary Volumes 19*, 119–150.

Karen Green
The University of Melbourne
Karen.green@unimelb.edu.au

Original Scientific Paper
UDC 17.024.1
165.194
Received: July 26, 2024
Revised: September 12, 2024
Accepted: October 18, 2024

THE MORAL PROBLEM IS A HUME PROBLEM

Abstract

The moral problem, as articulated by Smith, arises out of the attempt to introduce the experimental method of reasoning into moral subjects, developed by Hume. This paper returns to Locke's earlier attempt to provide an empirically adequate account of morality and the debate his attempt generated. It argues that the seeds of a more adequate, naturalistic account of the metaphysics and epistemology of morals than that developed by either Locke or Hume can already be found in aspects of Locke's *Essay* and in the defence of his views published by Catharine Trotter Cockburn. Locke and Cockburn find a natural, intrinsically moral, human disposition in our tendency to judge the moral good or evil of persons or actions in the light of their conformity with a moral law. It is constitutive of our nature as social beings that we are endowed 'with a moral sense or conscience, that approves of virtuous actions, and disapproves the contrary.' Moral laws are those prohibitions and obligations that benefit others and society as a whole. Thus, the question of natural, moral motivation is seen to be independent of the question of the objective grounds of moral truth. In virtue of our nature as social beings we are motivated to do what is approved of by other members of our society. Whether what is approved of by a society genuinely fosters the welfare of its members is an independent, *a posteriori* question that can only be answered through reasoned, empirically informed debate.

Keywords: Conscience · cognitivism · naturalism · natural law · counterfactuals

1. Introduction

The development of the methods of empirical science, during the seventeenth and eighteenth centuries, challenged widely accepted assumptions concerning knowledge of both the material and the moral universe. Belief in God's creation of a world governed by fundamentally moral laws, as taught by religion, spelled out by revelation, or intuited by innate reason, was undermined by the new science, based on observation and experiment. This set up an opposition between science and religion that continues to haunt society to this day. Descartes attempted to navigate the problem by distinguishing the realm of the material universe, the mechanical operations of which could be known

 DOI: 10.5937/BPA2437103G

by observation, from that of the immaterial mind, which remained knowable through introspection and innate reason. Locke in his *An Essay Concerning Human Understanding* more uncompromisingly attempted to demonstrate that the foundation of all human knowledge resides in experience, while, as we will see, also allowing a place for a kind of introspective experience. Whereas Descartes had retained innate knowledge of geometry, mathematics, and morals, Locke denied the existence of innate knowledge of any principles, whether of logic, mathematics, or morals. Hume's *Treatise of Human Nature*, likewise, proposed to extend the empirical methods of science to the moral realm and is subtitled 'An Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects' (Hume 1978). In what follows I argue that Hume's was a flawed attempt, condemned from the outset, in virtue of the way in which it built on the underlabourer, Locke's, unstable foundation. The moral problem that persists to this day, as set out by Michael Smith in various works, can be dissolved by repairing that flawed foundation (Smith 1994; 2009). At the same time, the seeds of a more adequate, naturalistic account of the metaphysics and epistemology of morals than that developed by either Locke, Hume, or Smith can be found in aspects of Locke's *Essay* that were overlooked by Hume.

To make this case I first examine the reception of Locke's own attempt to characterize our knowledge of moral truth, outlining the objections that were immediately raised against it. Next, I consider a repair to Locke's moral epistemology, developed by one of his defenders Catharine Trotter Cockburn. She proposed that morality is grounded in the fitness of things, developing a line of thought similar to that of Samuel Clarke. Hume's *Treatise* offers a direct response and critique of the resulting account of the nature of moral truth. This sets the scene for the moral problem that exercises Smith. Yet the problem only arises, in the form that it does, because Locke's flawed foundation is developed in a particular direction. Already there exists, in Locke's own writing, the materials for a more adequate, empirically based account of the nature of morality than that developed by Hume. This results in the dissolution of the moral problem, as set out by Smith.

2. Locke's moral epistemology and its critics

There is considerable confusion and controversy over the exact character of Locke's moral epistemology. It has been suggested that, depending on where one looks in the *Essay*, 'our knowledge of moral principles seems to depend on *a priori* reasoning, social learning, or the analysis of terms' (Wilson 2007, 381). As we will see this ambiguity results from the fact that different questions are being answered by Locke in various sections of the *Essay.* In the Introduction he is interested in the question of whether knowledge of moral principles is innate. Later his concern is to explain how we come by the idea of morality and what morality is, finally, he gives an account of the nature of our access to moral truth.

Soon after the publication of Locke's work an anonymous critic, who I will call 'the Remarker', published *Remarks upon an Essay Concerning Humane Understanding*, in which he accused Locke's empiricist epistemology of being unable to explain how we acquire knowledge of objective moral truth, or knowledge of the attributes of God, such as his goodness, or knowledge of the immortality of the soul (Watson 1989, 24). Knowledge of these, he argued, cannot be acquired by means of sensation or from reflection on ideas acquired through sensation.[1] With regard to moral truth, the upshot of his complaint is that the best an empiricist can do, by way of an account of the distinction between good and evil, is to conclude that the confection of such a distinction has been useful to society and government. But this is far from amounting to a demonstration that vice and virtue are grounded in intrinsic and immutable features of the nature of things. Using Michael Smith's vocabulary, the Remarker can be seen to be proposing that a consistent empiricist must really be a moral nihilist, since any naturalistic, utilitarian, or epicurean account of morality of the kind available to the empiricist cannot explain the existence of immutable, objective, moral truths (Smith 2009, 181–4).

In 1702, Locke was defended against these criticisms by Catharine Trotter Cockburn, who developed her own account of the nature of moral obligation and the objective grounds of moral truth, based on a clarification of what she understood to be Locke's position (Cockburn 1702; 1751; Bolton 1996; Sheridan 2007; 2018; Green 2019). Although Locke had rejected the view that humans possess innate knowledge of moral principles, arguing on the basis of observation, that different societies diverge markedly with regard to moral beliefs, he had also claimed that the idea of a Supreme Being, taken in conjunction with 'the *Idea* of ourselves, as understanding, rational Beings, being such as are clear in us, would, I suppose, if duly considered and pursued, afford such Foundations of our Duty and Rules of Action as might place *Morality amongst the Sciences capable of Demonstration* ... as incontestable as those in Mathematicks' (Locke 1975, IV.iii.18, 549). According to Locke our knowledge of morality is grounded in the recognition of relations among ideas, as is our knowledge of mathematics, but these are not ideas derived from sensation, but ideas derived from reflection on the idea of God and on our nature as rational beings. To modern readers, sensitive to the difficulties that attend an empiricist account of *a priori* truth in general, Locke's position looks heroic.

That it was heroic is confirmed by the repair that Cockburn later made to Locke's position in shoring up her defence of it. Locke claimed that reflection on our idea of ourselves as practical rational beings, plus the idea of God, will show us that moral truths are demonstrable––a surprisingly rationalist proposal. The thought that morality follows from our nature as

1 Since there is controversy over the identity of the author of the *Remarks*, it is safest to call him 'the Remarker', see (Walmsley et.al. 2006).

practical rational beings anticipates the anti-Humean and broadly Kantian, cognitivist account of morality, argued for by Smith (Kant 1990; Smith 1994, 130–81). Yet, unlike Smith, Locke assumes that objective morality depends on the existence of God as well as on practical rationality. Unlike Kant, Locke believes that God's existence can be proved from the idea of God, as derived from experience (Locke 1975, IV.x.1, 619). This supplies a reason to believe in the existence of a law maker, whose law Locke claims we are obliged to follow, in virtue of being his creation (Locke 1975, II.xxviii.4–8, 351–2).[2] Like Kant and Smith, he claims that the principles of this law can be known *a priori*. But the truths known depend for their existence on God the creator, as do mathematical and scientific truths.

In drawing the analogy between mathematics and morality Locke also assumes the possibility of an empiricist account of the *a priori* truth of mathematics, a possibility still hotly contested, since many are convinced that numbers are abstract objects and not part of the causal realm (Benacerraf 1973). The case for accepting an empiricism that makes moral truths knowable *a priori* is even weaker, since there is far less agreement over moral truth than over mathematical truth. Examples of *a priori* moral truths suggested by Locke, such as '*Where there is no Property there is no Injustice*' are simply unconvincing (Locke 1975, IV.iii.18, 549). This generalization is open to empirical refutation. Our moral ideas are derived from the moral practices of our society, and these are culturally variable. Societies in which injustice is recognised, but there is no property, are conceivable and may even have been actual. So, in defending Locke, Cockburn is faced with a problem, similar to that which Smith sets himself, that of providing a naturalistically acceptable account of our *a priori* knowledge of morality. Smith attempts to do this without assuming the existence of God. Moral judgments become 'expressions of our beliefs about what we have reason to do, where such reasons are in turn categorical requirements of rationality' (Smith 1994, 185). Cockburn also believes that moral obligations are requirements of rationality but, as we shall see, she repairs Locke's position by falling back on the existence of possible worlds, one of which, the actual world, God has chosen to instantiate. The nature of her repair suggests that no substantive solution to the problem of providing a naturalistic account of the cognitive grounds of morality, as Smith attempts to do, is available, unless God is surreptitiously reintroduced.

In the first version of her reply to the Remarker, Cockburn had proposed that the grounds of morality lie in our human nature. Here she was following Locke who, although he had denied that humans possess innate knowledge of moral principles, had said that those who 'deny that there is a Law knowable

2 Locke's claim that law requires a law maker has led some to deem him a voluntarist, (Darwall 1995, 37; Colman, 1983, 5; Sheridan, 2007, 143). That Cockburn is right to read him as an intellectualist is argued in Green, 2019).

by the light of Nature' fall equally into error with those who assume innateness (Locke 1975, I.iii.13, 71). When her reply was republished in her collected works, Cockburn added two clarificatory footnotes to it, which emphasized that relations among ideas are not simply relations among ideas as they are in the mind, but relations among ideas as they are in the mind of God, who has chosen to create things with the corresponding natures (Cockburn 1751, I.56, note f, 62, note k). Her first footnote made clear that insofar as '*the nature of man is the ground or reason of the law of nature; i.e.* of moral good and evil' there must be a real and everlasting truth as to the nature of man (I.57). Now, following Samuel Clarke, she accepts that this leads 'us to the supreme mind, where all truth, and the abstract nature of all possible things, must eternally and immutably exist' (I.56, note f). In the second footnote she says that the law of reason and the law of nature oblige us as *reasonable* beings 'in the same manner as the Supreme Being, who is subject to no laws, and accountable to none, obliges himself to do always what he perceives to be right and fit to be done' (I.62, note k).

The result is a theistic, realist, naturalism that in many ways harks back to Stoicism (Sheridan 2018). It relies on there being a truth concerning human nature, and certain acts being fit or appropriate for humans, given that nature. It is an account of moral obligation that can be recognised as being in Hume's sights when he says in the *Treatise of Human Nature*,

> Those who affirm that virtue is nothing but a conformity to reason; that there are eternal fitnesses and unfitnesses of things, which are the same to every rational being that considers them; that the immutable measure of right and wrong impose an obligation, not only on human creatures, but also on the Deity himself: all these systems concur in the opinion, that morality, like truth, is discerned merely by ideas, and by their juxtaposition and comparison. In order therefore, to judge of these systems, we need only consider whether it be possible from reason alone, to distinguish betwixt moral good and evil, or whether there must concur some other principles to enable us to make that distinction. (Hume 1978, III.1.§1)

From here he sets up the moral problem set out by Smith. Beliefs, Hume claims, do not motivate.[3] If morality were simply grounded on relations among ideas, as is mathematics, then we would be no more motivated to act on the judgement that the unjustified killing of an innocent is murder, than

3 The theistic naturalists would disagree. They adopt the principle of moral necessity. It is impossible for a completely good agent to believe that an act is the best possible act and not be motivated to do it. This is the reasoning that leads Leibniz to conclude that God must have created the best of all possible worlds. Interestingly Cockburn gives God more freedom, by assuming that there might be equally good possible worlds, (Thomas 2017). Theistic naturalism fails because it is impossible to prove, on the basis of ideas derived from sense, that there exists an infinitely good, supreme being.

we are by the judgement that the square root of 81 is nine. Moral judgements motivate, so cannot be solely grounded in relations among ideas.

Hume's own response to this problem is, in effect, to adopt a version of the naturalist position that the Remarker claimed was the only one available to a consistent empiricist. We approve of moral actions and virtues, in ourselves and in others, because we have come to believe that such actions and virtues promote the natural physical pleasures we desire, and they tend to reduce the natural physical pains we wish to avoid. G. E. Moore's open question argument, however, suggests a conclusion congenial to the Remarker. No such reduction of morality to a means of maximizing natural value can be truly adequate, since it always makes sense to question whether an act that maximizes the satisfaction of physical pleasure is actually morally good. Though he gives it an important place in the history of attempts to address the moral problem, Smith does not accept this rejection of naturalism, because, he argues, it fails to distinguish motivating from rationalizing reasons. But another response is to object that G.E. More assumes that there are no natural motivations that are intrinsically moral. It is here that an overlooked aspect of Locke's *Essay* is relevant.

3. An alternative strand in Locke's account of morality.

In later writings, Cockburn sets out her own understanding of morality as grounded in natural law and says,

> Mankind is a system of creatures, that continually need one another's assistance, without which they could not long subsist. It is therefore necessary, that every one, according to his capacity and station, should contribute his part towards the good and preservation of the whole, and avoid whatever may be detrimental to it. For this end they are made capable of acquiring social or benevolent affections, (probably have the seeds of them implanted in their nature) with a moral sense or conscience, that approves of virtuous actions, and disapproves the contrary. This plainly shews them, that virtue is the law of their nature, and that it must be their duty to observe it, from whence arises *moral obligation* ... (Cockburn 1751, I.413)

We can translate this into two propositions; first that humans are social animals, second, that in virtue of being social animals, humans can acquire social affections and have 'a moral sense or conscience, that approves of virtuous actions, and disapproves the contrary.' In more modern language, Cockburn is saying that it is constitutive of the fact that humans are social animals that they are possessed of social affections and a conscience. Hume also allows that there are natural social affections, but he fails to consider the

possibility, here raised by Cockburn, following Locke, that the possession of a conscience might also be part of human nature. For, while Locke denied that knowledge of moral principles is innate, he did not deny that humans have an innate tendency to judge their own actions in the light whatever moral law is established by the customs of their society (Watson 1989, 67–8).

In book two of the *Essay*, Locke set out to show how we come by all our ideas, either from sensation or by reflecting on the operations of our own minds. Both physical and mental pleasures and pains are natural. We derive ideas of the passions such as love, fear, envy, and shame, when we reflect on ourselves, and observe how both pleasure and pain, and the good and evil which cause pleasure and pain operate in us and 'what modifications or tempers of mind, what internal sensations (if I may so call them) they produce in us' (Locke 1975, II.xx.3). Among the passions are love, which derives from 'the delight that some present or absent thing is apt to produce' and shame, 'an uneasiness of the mind upon having done something which is indecent, or which will lessen the valued esteem which others have for us' (II.xx.4 & 17). Although he does not discuss the matter, nothing that Locke says implies that such passions are not natural. The idea of shame is derived from reflecting on feelings of the mind and is closely related to conscience. In her response to the Remarker, Cockburn had commented that he had not sufficiently considered that Locke had included ideas of reflection, as well as ideas gained from sensation, as underpinning moral obligation. If social and moral sentiments such as love and shame are natural to us, the naturalist can avail herself of these much richer materials in developing a metaphysics and epistemology of morality.

Later in book II, Locke discusses the idea of morality as a relational idea. Moral good and evil, he says, 'is the conformity or disagreement of our voluntary actions to some law' (II.xxviii.5). The law in question may be divine, civil, or that of opinion or reputation (II.xxviii.7). '"Virtue" and "vice" are names pretended and supposed everywhere to stand for actions in their own nature right and wrong' (II.xxviii.10). In so far as actions conform or fail to conform to the divine law, they are in fact virtuous or vicious. But he allows that, throughout the world, they are attributed to actions that are approved or disapproved of.

> Thus the measure of what is everywhere called and esteemed virtue and vice is the approbation or dislike, praise or blame, which by a secret and tacit consent, establishes itself in the several societies, tribes and clubs of men in the world, whereby several actions come to find credit or disgrace amongst them according to the judgement, maxims, or fashions of that place. (II.xxviii.10)

Different societies have different standards of what is right and wrong. But everywhere people judge the virtue or vice of their actions in relation to whatever the law is that sets the standard for their society. Here Locke is

pointing to facts that have induced others to move towards cultural relativism. He does not accept such relativism, because he believes that, as well as civil and conventional law, there is divine law. Just as many cultures are ignorant of scientific facts, so they may mistake the divine law. As we have seen he believes that the idea of a Supreme being, taken in conjunction with 'the *Idea* of ourselves, as understanding, rational Beings, being such as are clear in us,' can result in knowledge of the divine law, thus giving us access to the true nature of virtue and vice (IV.iii.18, 549).

Locke's path to natural but intrinsically moral motivations is not considered by Hume. He begins his discussion of the passions by following Locke in distinguishing original or sensory impressions from secondary or reflective impressions which he calls passions (Hume 1978, II.1. §1). Yet he never considers the possibility that it might be part of our nature, as social animals, to have a disposition to judge the appropriateness of our actions in relation to some law. Rather, he reduces the motivational aspect of the passions to the pains and pleasures that they produce in us according to the association of ideas. The thought that moral dispositions might be both natural and grounded in our nature as self-conscious reasoning beings is overlooked by him.

Superficially, the theological, naturalist realism, implicit in Locke and spelled out by Cockburn, offers little comfort to the modern naturalist. Objective moral truth is purchased at the cost of accepting that ideas of human nature exist eternally in the mind of God. These are surely 'queer' entities of the kind pointed to by error theorists such as John Mackie (Mackie 1977). We appear to have fallen back into nihilism or at best cultural relativism. Nevertheless, the account that Cockburn builds, on the material provided by Locke, also points to a different path to an internalist naturalism to that taken by Smith. Rather than being completely anti-relativist, as is Smith's more Kantian position, it arguably retains what is right in cultural relativism without abandoning objectivity.

4. Dissolving the moral problem

The moral problem that exercises Smith consists in three plausible but incompatible propositions.

1. Moral judgments are about objective matters of fact.
2. Moral judgments motivate.
3. Motivation is Humean, it depends on desires and means-end belief (Smith 1994, 12).

Smith solves the problem by rejecting Humean motivation; rationalising reasons motivate as well as desires (Smith 147–81). The solution built on materials derived from Locke and Cockburn agrees that reason motivates

rational individuals, but also proposes that conscience, interpreted as the desire for self-approbation, is a fundamental moral motivation, so the apparent incompatibility dissolves. Recognising conscience accounts for the ubiquitous conflict between the demands of morality and what self-interested reason may suggest. Objectivity is retained, since there are objective matters of fact as to what the moral laws require and these laws are themselves open to objective evaluation, as solutions to problems of social co-ordination. This implies a certain relativism but explains the grounds of reasoned moral debate without relying on the possible worlds to which, it will be argued, Smith must ultimately appeal.

Smith claims that our concept of moral rightness is the concept of what we would desire ourselves to do were we fully rational, where the substantive content of the desire involves contributing to human flourishing (Smith 1994, 184–5). This is like Cockburn's position. The content or aim of morality is the social good and we have reason to act on our beliefs as to what will bring about that good. What will bring about the good, according to her, is acting in a way that is fit, given our God given nature. Unlike her, Smith does not explicitly assume that we have a God given nature. However, there being a truth as to what we would desire, were we fully rational, requires the truth of a counterfactual. There are significant questions to be raised concerning the truth conditions of counterfactuals (Dummett 1993, 248–54). Those who believe that they can be true, explain their truth conditions in terms of possible worlds (Lewis 1973). Cockburn was forced to introduce ideas in the mind of God, that determine what is possible, to shore up her defence of Locke. Smith and Cockburn then, end up being committed to the same queer entities–– possible worlds––as truth makers for *a priori* moral truths.[4]

The thread to guide us along the path to a sufficiently objectivist naturalism is the observation that Locke and Cockburn do not actually believe that 'virtue is nothing but a conformity to reason.' True virtue, they claim, is conformity to reason, based on knowledge of human nature, but what passes for virtue, in most societies, is conformity with whatever is believed good, according to customary moral law. As social beings we are endowed 'with a moral sense or conscience, that approves of virtuous actions, and disapproves the contrary' (Cockburn 1751, I.413). But this 'moral sense' should not be thought of as a means of perceiving moral truths. Rather it is a natural disposition to judge our own actions in the light of their conformity or lack of conformity to established social custom. Unless custom can be shown to conform to a God given standard, this moral truth will not be

4 Like Locke, Smith thinks that moral knowledge is *a priori*. At least he says that 'it is a relatively a priori matter,' (2009, 203). He does not seem to recognise, as does Cockburn, that this means that he owes us an account of the existence of objective ideas as to human nature. Possibly, his admission that his own non-relativistic, internalist, naturalist, moral realism may rest on an illusion is an admission that, in the end, it presupposes the existence of God (205).

completely objective. If there is no God given standard, moral laws may fail to promote the good and so be rationally criticisable but not absolutely objective. Objective moral truth requires an objective, non-relative truth as to human nature, and this, it will be argued, is not available for partly self-constituting, ethical beings like us.

According to Locke, what counts as virtuous action is that which receives approbation, as a result of being in conformity with whatever is established, in a society, as being morally good. What is morally good or bad is simply what the people morally approve and disapprove of. Being approved or disapproved of by others, motivates conformity to whatever law is established. For,

> He who imagines commendation and disgrace not to be strong motives on men to accommodate themselves to the opinions and rules of those with whom they converse, seems little skilled in the nature or history of mankind; the greatest part wherof we shall find to govern themselves chiefly, if not solely, by this *law of fashion*; and, so they do that which keeps them in reputation with their company, little regard the laws of God, or the magistrate. (Locke 1975, II.xxviii.12, 356–7)

This offers a naturalistic theory of moral motivation that grounds it in an innate desire to be approved of by members of one's society. Being approved by members of one's society results in those feelings of self-approbation that accompany the judgment of oneself as being a good person. Being disapproved of is pain to (most) social animals, like humans, and results in feelings of guilt, or shame, reflective passions that are only available to creatures that are able to assess their own actions, in the light of the socially sanctioned standards of behaviour that we call moral laws.[5] Conscience is, in Smith's language a motivating reason. What Cockburn means by conscience refers back to this idea of morality, for she says it 'is nothing else but a Judgement which we make of our Actions, with reference to some Law, which we are persuaded ought to be the Rule of them' (Cockburn 1702, 71). Locke also uses the term 'conscience' in this sense. Conscience is not an innate vehicle that delivers knowledge of morality, but it is an innate disposition to judge one's own actions in the light of whatever morality is established in one's community.

A form of moral naturalism based on Locke's account is not subject to critique via the open question argument.[6] It is always an open question as to whether the moral laws established by a society maximize natural physical pleasure. It is not an open question whether the moral laws of a society are among those rules that members of the society conform to, in order to gain the approbation of other members of the society. Moral sentiments are

---

5 The sociopath is someone who, for whatever reason, has failed to acquire such normal moral motives.

6 That the open question argument is not sound is now accepted by Smith, (2009, 200).

genuinely moral, on this view. But the position is open to the objection that it deprives morality of objectivity. On this account, how is morality different from politeness or fashion? We are back to cultural relativity. To add a measure of moral realism, we need to recognise that the account of moral motivation needs to be supplemented by a theory of moral judgement that explains both how judgements internal to a system of law can be objective and how some systems of moral laws can be judged to be more adequate than others. We need an account that distinguishes moral principles from those of fashion, taste, or beauty. That is to say, in Smith's language, we need an account of rationalising moral reasons. Before developing this, the view needs to be defended from the objection that what has been offered is not an account of genuinely *moral* motivation.

Cockburn recognises three sources of motivation, deriving from the fact that humans are rational, social, and sensible beings (Cockburn 1751, I.420; Sheridan 2007, 254). As sensible creatures we wish for pleasure and to avoid pain. As sociable creatures, we are disposed to care about the welfare and judgement of others. As rational creatures we desire to act in accord with what reason tells us to be the case, 'to act contrary to the reason, relations, and fitness of things,' she says, 'may not improperly be called the *pain* of rational being' (I.420). It is a significant aspect of her worked out view that all three sources of motivation are natural to us as humans. We have social affections and are rational as well as sensible beings. Implicit in her view is the acceptance that what distinguishes moral laws is that they are those that relate to behaviour that impacts on the welfare of others, and they are rationalizing in so far as they are grounded in knowledge of the truths of our nature.

What Cockburn says about social affections may well have been influenced by Hutcheson, for both mention the natural disposition that parents have to care for their children. He speaks of the honest farmer who, 'studies the preservation and happiness of his children without any design of good to himself' (Hutcheson 2008, 112; Harris 2015, 70). She asks, 'Can any one think, that the fondness of a mother, and her tender concern for the happiness of her child, is owing to her "having perceived, or been taught from her infancy, that her happiness is necessarily connected with that of others; that their esteem is useful to her ..."' (Cockburn 1751, I.427). She clearly has in her sights those, like Bernard Mandeville, who had attempted to explain such social affections as grounded in self-interest. Like Hutcheson she insists that they are genuinely other oriented. But since Hume had also rejected Mandeville's kind of psychological egoism, one might wonder whether bringing in the social affections can really avoid the open question argument. Might it still not be an open question whether a system of laws that maximizes the satisfaction of all the physical and social desires of a people, including their desires to be approved by others, is a genuinely moral system?

Locke would answer 'yes' to this question. People who are not enlightened by knowledge of the deity, might have a well-functioning set of

moral laws, and be virtuous in the sense of following their consciences and acting in accord with established law, but still be mistaken as to the moral truth. On this view, the idea of morality might be better compared with the idea of medicine than with mathematics. Medicine is the science of healing bodies. Different societies have different bodies of medical belief. There is ancient Chinese medicine, Gallenic medicine, twentieth-century Western medicine. These are 'medicines' thought of as social practices. But there is also 'true' medicine, that which tracks the actual functioning of the human body. 'Morality' likewise is a social practice. If there is to be a 'true' morality, there must be, as Cockburn concluded, a moral truth grounded in the truth about human nature. This is what Locke says is promulgated 'by the light of Nature or the voice of Revelation' (Locke 1975, II.xxviii.8, 352). It cannot be found in the mere ideas of human nature that different human societies have developed, that is to say, ideas as they exist in the human mind. Cockburn concludes that it must reside immutably and internally in the mind of God. I do not believe that this kind of theistic, immutable, moral truth is available to a genuine naturalist. The consistent empiricist will have to accept that, in one sense of 'objective moral truth', moral truths are social truths. What the moral laws of a society are at a time is a reasonably objective, social fact, but the empiricist does not have to accept that this leads back to nihilism. There is another kind of objectivity available, that is grounded in the purpose of the social practice. Moral laws are those that impose prohibitions and obligations on individuals that are taken to be necessary for the welfare of other individuals and the welfare of the society in general. So actual moral laws will be somewhat relative. They will be moral in virtue of their function, which is the overall welfare of the people who obey them. Yet they will be relative to the means available to secure that welfare and, potentially, different conceptions of welfare. The position is therefore somewhat like that developed by Gilbert Harman or David Gauthier (Harman 1975; Gauthier, 1987). For Smith this is not sufficient. Relativism, he claims, is not consistent with objectivity. He attempts to develop a non-relativistic, internalist, naturalist, moral realism but he fails to show that this does not, in the end, presuppose theism.

It is here that a fundamental fissure in the foundations of the Lockean account of knowledge needs to be recognised. Locke had begun his *Essay* by claiming that the immediate objects of perception are ideas and he only later distinguished ideas 'as they are in our minds' and as they are dispositions in things to affect us (Chappell, 1994). Locke's 'ideas' are sometimes sensations caused in us by powers in things, at other times they are the powers in things themselves. Sensory ideas, the effects of those powers that impinge on us through our senses are adequate in giving us real knowledge of things, though not of the fundamental causal mechanisms that underlie the powers (Locke 1975, II.xxxi.2, 375–6). Hume distinguishes impressions, thought of as sensations, from ideas in the mind. His 'ideas' are purely 'in the mind', resulting in scepticism with regard to the existence of mind-independent

reality and with regard to causal relations, thought of as involving necessary connection. If ideas are 'in the mind' then it appears that we cannot fall back on truths concerning human social nature to underpin our inquiry into objective morality. But if there are genuinely human social dispositions, a naturalist can avail herself of the existence of such social dispositions to underpin objective moral science. Just as there are truths about human physical nature, that can underpin objective medical science, there may be truths about humans as social beings, that can underpin objective moral science. And it is, as Locke insists, a fact about humans that commendation and disgrace are strong human motivators.

Smith attempts to defend moral objectivity by developing a version of the internalist option pursued by Kant. The passions are not the only source of human motivation, reason also motivates. Reason, Kant had claimed, desires to guide itself by a rational law (Kant 2018). Smith similarly finds moral motivation in the desire that a human has to choose what they would desire were they more perfectly rational and well informed than they actually are. Both are guided by Hume's subjectivism into thinking of the moral problem as a problem of individual motivation. They fail to recognise the other internal source of moral motivation, as recognised by Locke and Cockburn. It is an empirical fact that our nature as social beings means that we are innately disposed to judge our own actions, in the light of the moral laws established in our society. Reason is also a source of internal motivation. We do desire to determine our actions on the basis of true beliefs as to the nature of things. The rational medical practitioner does not merely accept the medical beliefs of their society, he or she subjects them to rational inquiry to determine whether they actually promote health. Equally, the rational moral practitioner does not merely accept the moral beliefs of their society, he or she subjects them to rational inquiry to determine whether they actually promote human welfare. The clash that can arise between the two sources of moral motivation, what society requires and what a more expansive reason suggests it ought to require, is the stuff of much great literature.

From this point of view, the moral problem that exercises Smith dissolves and is replaced by a different more urgent moral problem. This is the question of which socially sanctioned rules of behaviour will promote human welfare in the environmental, historical, and technological circumstances in which we now find ourselves. People are strongly motivated to mimic the behaviour, to seek the approval, and to avoid the disapprobation of others. This is the engine that fuels social learning of both linguistic and moral conventions. Yet this leads to conflict with those who have learned to conform to different ways of being. In both language and morality there is a certain arbitrariness and a certain lack of arbitrariness. The conventions of language work because they are shared by a population and the information conveyed by the language is adequate to the needs and pursuits of the people speaking

it. Many different systems of linguistic conventions are adequate, so long as their speakers conform to the standards that are necessary for language to perform the functions for which it evolved. Languages need to be sufficiently good vehicles for conveying information. They need to convey information about the things that exist in the environment in which the people operate. Language users need to be able to convey clear and consistent messages. They need to be sufficiently trustworthy sources of information, for the language to be of benefit. The evolution of language brings with it moral injunctions not to lie to those who are trustworthy (friends), and not to convey certain kinds of information to others who are not trustworthy (enemies). Language brings with it, norms of language use. Equally, without language, moral conventions and moral motivation could not have got off the ground. The articulation of a publicly recognised moral law is impossible without language. The development of a self-conscious sense of the self, as approved or disapproved of by others, depends on people possessing the capacity to convey their judgements of their own and other's conformity, or lack of conformity to a law, as articulated in language. Conscience is closely related to the kind of self-consciousness that is associated with the capacity to think about the beliefs of others, as conveyed in language.

This destroys the sharp fact value distinction that Hume introduced and the moral problem that results from it. Our theories of what we are, our understanding of our nature, bleeds into our understanding of how we ought to behave. We are by nature social creatures, whose sociability manifests itself in language and in the desire to conform to the law of our nature, operative in the social group into which we have been born. The best current account of our nature is that language, along with the capacity for reasoning, self-consciousness, and morality, evolved because they gave our species an evolutionary advantage. The importance to us of a sense of self-worth, that is tied to the conventional morality of our group, is both a strength and a weakness. For while it usually fosters in-group co-operation it also often motivates inter group conflict, particularly when groups that have evolved different moral codes come into contact with each other, or when well established codes cease to benefit overall welfare, because of environmental or technological change.

Smith argues that context relativity is not compatible with objectivity. But relativity is not necessarily at odds with objectivity. In physics, the time it takes for an object to travel a distance is relative to the speed at which the observer is traveling. This is objective but relative. Which moral principles will foster the well-being of a society are plausibly relative to the available technology, scientific knowledge, and environment that form the background of non-moral facts within which the society operates. On this view, moral truths are neither immutable nor are they completely arbitrary. They are social truths about the laws against which various people judge the

appropriateness of their behaviour. This is one source of the internalist aspect of moral vocabulary. There are also truths about the effectiveness of various systems of moral principles in achieving the goal that morality evolved to achieve. That is, promoting the welfare and survival of the species in which it evolved. This accords with the views of externalist realists (Smith 2009, 201). Just as the norms of language arise from its purpose as a vehicle for reasoning and communication, so to, the norms of morality arise from its function as a means of social co-ordination and co-operation. In both cases, the element of arbitrariness means that many different conventional systems may function equally well. This does not detract from the fact that others may have been rendered dysfunctional by changing circumstances.

The choice between Smith's Kantian cognitivism and the proposed modification of Lockean cognitivism comes down, then, to the choice between an implicitly theistic position and a somewhat relativistic one. For if there is to be a truth as to what one would do, were one fully rational, there must be a truth as to what is possible. There must be a truth concerning human nature and what is morally required given that nature. Kant had become convinced that pure speculative reason was incapable of proving the existence of God. The natural law theorists were therefore unable to ground the objectivity of morality. He argued instead that belief in the possibility of a good will, that freely chooses to be guided by a rational moral law, presupposes the existence of God. It is then from the presuppositions of pure practical reason that freedom, God, and immortality can be taken to derive "Bestand" (standing) and "objektive Realität" (objective reality) (Kant 1910–, 5.3–4; 2015, 3). If there is a truth as to what one would do, were one fully rational, there must be a God whose existence underpins that moral truth.

The consistent naturalist cannot concur that human nature is fixed by God, we are social creatures who are, up to a point, self-constituting. What is best for us is not determined by natural or by God given facts. Hence, moral principles are inevitably contextual, but that does not imply that they are not objective in a relevant sense. Moral laws are conventions the purpose of which is to foster social co-operation. They should be made fit for the circumstances in which humans find themselves. Principles that arose because adapted to the needs of relatively isolated agricultural communities are not necessarily adaptive for individuals living in a global, industrialised world. Looked at from this point of view the pressing moral problem is not, 'What ought I to do?' I ought to obey the moral law that I believe people in the narrower and broader communities, within which I operate, ought generally to obey. The pressing problem of our times is how to come to an agreement over the content of such moral laws, in the circumstances in which we now find ourselves. The pressing moral problem is how to reach agreement on moral principles, the implementation of which will actually foster the kind of co-operation necessary to preserve the existence of our species, thus achieving

the purpose for which morality evolved. The solution to this problem cannot be known *a priori* but has to be grounded in significant *a posteriori* investigation, negotiation, and a will to foster co-operation.

References

Benacerraf, Paul. 1973. "Mathematical Truth," *Journal of Philosophy* 70, no. 19; 661–79.

Bolton, Martha Brandt. 1996. "Some aspects of the philosophical work of Catharine Trotter Cockburn." In *Hypatia's Daughters: Fifteen hundred years of women philosophers*, edited by Linda Lopez McAlister, 139–64. Bloomington: Indiana University Press.

Chappell, Vere. 1994. "Locke's Theory of Ideas." In *The Cambridge Companion to Locke*, edited by Vere Chappell, 26–55. Cambridge: Cambridge University Press.

Cockburn, Catharine Trotter. 1702. *A Defence of the Essay of Human Understanding Written by Mr Lock. Wherein its Principles with reference to Morality, Reveal'd Religion, and the Immortality of the Soul, are Consider'd and Justify'd: In Answer to Some Remarks on that Essay.* London: Printer for Will Turner at Lincolns-Inn Back-Gate, and John Nutt near Stationers-Hall.

Cockburn, Catharine Trotter. 1751. *The Works of Mrs. Catharine Cockburn, Theological, Moral, Dramatical and Poetical,* edited by Thomas Birch. 2 vols. London: J. and P. Knapton.

Colman, John. 1983. *John Locke's Moral Philosophy*. Edinburgh: Edinburgh University Press.

Darwall, Stephen. 1995. *The British Moralists and the Internal 'Ought': 1640–1740*. Cambridge: Cambridge University Press.

Dummett, Michael. 1993. "Realism," in *The Seas of Language*, 230–76. Oxford: Oxford University Press, 1993.

Gauthier, David. 1987. *Morals by agreement*. Oxford: Clarendon Press.

Green, Karen. 2019. "On some footnotes to Catharine Trotter Cockburn's *Defence of the Essay Of Human Understanding*." *British Journal for the History of Philosophy* 47, no. 4: 824–41.

Harman, Gilbert. 1975. "Moral Relativism Defended," *The Philosophical Review* 84, no. 1; 3–22.

Harris, James A. 2015. *Hume. An Intellectual Biography.* Cambridge: Cambridge University Press.

Hume, David. 1978. *A Treatise of Human Nature.* Oxford: Clarendon Press.

Hutcheson, Francis. 2008. *An Inquiry into the Original of our Ideas of Beauty and Virtue.* 2nd ed. Indianappolis: Liberty Fund.

Kant, Immanuel. 2015. *Critique of Practical Reason*, trans. Mary Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel. 2018. *Groundwork for the Metaphysics of Morals.* Trans. Allen W. Wood. New Haven: Yale University Press.

Kant, Immanuel.1910–. *Kritik der Praktischen Vernunft* in *Kant's gessammelte Schriften*. Berlin: Georg Reimer.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

Locke, John. 1975. *An Essay Concerning Human Understanding,* edited by Peter H. Nidditch. Oxford: Clarendon Press.

Mackie, John. 1977. *Ethics: Inventing Right and Wrong.* Harmondsworth: Penguin.

Sheridan, Patricia. 2007. "Reflection, Nature, and Moral Law: The Extent of Catharine Cockburn's Lockeanism in her *Defence of Mr. Locke's Essay*." *Hypatia* 22, no. 3: 133–51.

Sheridan, Patricia. 2018. "On Catharine Trotter Cockburn's Metaphysics of Morality," in *Early Modern Women on Metaphysics*, edited by Emily Thomas, 247–65. Cambridge: Cambridge University Press.

Smith, Michael. 1994. *The Moral Problem.* Oxford: Blackwell.

Smith, Michael. 2009. *Ethics and the a priori.* Cambridge: Cambridge University Press.

Thomas, Emily. 2017. "Creation, Divine Freedom, and Catharine Cockburn," in *Women and Liberty, 1600–1800*, edited by Jacqueline Broad and Karen Detlefsen, 206–20. Oxford: Oxford University Press.

Walmsley, J. C., Hugh Craig, and John Burrows. 2016. "Authorship of Remarks Upon Locke's Essay Concerning Human Understanding." *Eighteenth-Century Thought* 6: 205–43.

Watson, George (ed). 1989. *Remarks on John Locke by Thomas Burnet with Locke's Replies*. Doncaster, S. Yorkshire: Brynmill.

Wilson, Catherine. 2007. "The Moral Epistemology of Locke's Essay," in *The Cambridge Companion to Locke's "Essay Concerning Human Understanding,"* edited by Lex Newman, 381–405. Cambridge: Cambridge University Press.

Jussi Suikkanen
University of Birmingham
j.v.suikkanen@bham.ac.uk

Original Scientific Paper
UDC 17.022.1
17.036
Received: July 06, 2024
Revised: September 12, 2024
Accepted: September 17, 2024

HOW TO DISSOLVE THE MORAL PROBLEM

Abstract

According to Michael Smith, there is so much metaethical disagreement because it is difficult to explain both the objectivity and the practicality of moral judgments in the framework provided by the Humean picture of human Psychology. Smith himself hoped to solve this problem by analysing the content of our moral judgments in terms of what our fully rational versions would want us to do. This paper first explains why this solution to the moral problem remains problematic and why we therefore are no closer to solving the problem. It then outlines how the moral problem could perhaps be dissolved instead. The second half of the paper thus first reconstructs the moral problem in the framework of dispositionalism about belief. It then suggests that, if we think of moral beliefs in dispositionalist terms and take 'believe' to be a vague predicate, we can come to see why many of the most fundamental metaethical questions cannot be answered. The last section of the paper then extends this method of dissolving metaethical questions to other popular views about belief.

Keywords: Belief · Dispositionalism · Michael Smith · Moral Problem · Metaethics

1. Introduction – the Moral Problem

One remarkable thing about Michael Smith's 1994 book *The Moral Problem* is the big picture of metaethics, which its Chapter 1 provided.[1] After drawing the distinction between first-order moral and higher-order metaethical questions, Smith (3–4) pointed out that metaethicists disagree about almost everything. There is disagreement, for example, about whether moral facts exist, whether they are ordinary natural facts or *sui generis*, whether moral properties are causally efficacious, whether there is a necessary connection between moral judgments and motivation, whether moral judgments are beliefs or desire-like attitudes, whether moral requirements are requirements of rationality, and about whether morality is objective. Today, thirty years later, metaethicists still continue to disagree about these questions.

Smith, however, also wanted to explain why there is so much disagreement in metaethics. This explanation gives *The Moral Problem* its

1 Hereafter, all unattributed references are to Smith (1994).

 DOI: 10.5937/BPA2437121S

name. The problem is, according to Smith (5–7), that two features of our moral practices pull in different directions when we assume a Humean view of human psychology. The first of these features Smith called the 'objectivity of moral judgments' (6), which he captured in the following way (12):

1. Moral judgments of the form 'It is right that I φ' express a subject's beliefs about an objective matter of fact, a fact about what is right for her to do.

This feature of morality is the claim that, in moral inquiry, we are concerned about getting the answers to different moral questions right, which assumes that there are objectively correct answers to be had.

Smith called the second central feature of morality the 'practicality of moral judgment' (7), which he captured thus (12):

2. If someone judges that it is right that she φs then, *ceteris paribus*, she is motivated to φ.

This feature is based on the observation that, all else being equal, we expect people who make sincere moral judgments to be motivated accordingly. As Smith (7) put it, 'moral judgments seem to be ... opinions about reasons we have for behaving in certain ways, and ... having such opinions is a matter of finding ourselves with a corresponding motivation.'

The problem, according to Smith, is that these two features of morality have exactly the opposite implications in metaethical moral psychology when we assume the Humean picture of human psychology. According to that picture (7), there are two fundamentally different kinds of mental states: beliefs that purport to represent how the world is and desires that represent how the world is to be. On this view, beliefs are motivationally inert, but they can be evaluated in terms of truth and falsity. By contrast, desires are not assessable in terms of truth or falsehood as they are states of being motivated. Smith formulated the central crux of this psychological picture as follows (12):

3. An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume's terms, distinct existences.

The moral problem (i.e., the explanation of why there is so much disagreement in metaethics) then is that the propositions 1–3 form an inconsistent triad (12). 1 entails that moral judgments are beliefs, and 2 that they are necessarily connected to being motivated and hence according to 3 to desires. Yet, 3 states that no belief can have a necessary connection to a desire – believing that things are thus and so is one thing, and desiring the world to be in some way is something else.

This big picture also enables us to map the logical space of different metaethical views, and this map has been hugely influential – it has guided

a whole generation of metaethicists through the field. We can understand different metaethical views as rejections of one of the previous propositions 1–3. The expressivists and other non-cognitivists reject 1, the idea that moral judgments express beliefs; the externalist cognitivists reject 2, the necessary connection between moral judgments and motivation; and the anti-Humeans 3, the idea that beliefs and desires are distinct existences.[2]

Smith's (13) diagnosis of the state of metaethics in 1994 was that the disagreements between these positions will not go away because each position is trying to reject and explain away a proposition which seems more certain than the key elements of those positions themselves. The fact that the same metaethical disagreements continue to this day seems to confirm this diagnosis. Yet, after both diagnosing what is behind the fundamental metaethical disagreements and providing a map of the logical space, Smith also wanted to *solve* the moral problem. He argued that the alleged contradiction generated by 1–3 is merely an apparent one. Smith suggested that once we analysed the non-obvious content of our moral beliefs correctly, we would be able to explain the necessary connection between moral beliefs and motivation in rational agents within the framework of the Humean belief-desire psychology.

This is where this paper comes in. §2 first outlines an objection to Smith's solution, which has to do with its inability explain why certain combinations of beliefs and desires are incoherent. The rest of the paper outlines a way in which the moral problem could be dissolved rather than solved. For this purpose, §3 uses dispositional approaches to belief to construct an alternative big picture of the metaethical landscape. Just like Smith's big picture, it too will enable us to map the logical space of different metaethical views. §4 will, however, suggest that the adoption of this new way of seeing the metaethical landscape has significant consequences. Instead of offering a positive solution to the metaethical problems, it will provide us with a way of dissolving many of the central metaethical questions. Finally, §5 concludes by explaining why the proposed dissolution of the moral problem does not depend on the dispositional account of beliefs, but rather it can also be adapted to fit the frameworks provided by the other leading approaches to belief.

2. An Objection to Smith's Solution

As already mentioned, Smith (13–14) wanted to provide an analysis of the non-obvious content of our moral judgments to explain, within the Humean framework, both the objectivity and the practicality of those judgments. This analysis proceeded in two steps. Firstly, Smith (§3.2 and §3.6–§3.9) argued that moral judgments are judgments about what we have reason to

2 Smith (12–13) takes Ayer, Hare, Blackburn, and Gibbard to be proponents of the first strategy; Frankena, Foot, Scanlon, Railton, and Brink proponents of the second strategy; and Nagel, McDowell, Platts, McNaughton, and Dancy proponents of the third strategy.

do. Secondly, he (§5.9) also suggested that a judgment about what you have reason to do in a given situation is a judgment about what our fully rational versions would desire from their idealised perspective real people like you to do in that concrete situation in the actual world. The consequence of this analysis is that a moral judgment of the form 'It is right that I φ in C' would just be the belief that our fully rational selves in the evaluating world would want our non-idealized selves to φ in C in the evaluated actual world.

Let us then see how this analysis is supposed to solve the moral problem. Firstly, it seems to secure the objectivity of moral judgments in two ways. Firstly and more obviously, because moral judgments are on this view beliefs, they are be truth evaluable (185). There is some fact of the matter what desires we would have if we had all the relevant true beliefs, no false ones and deliberated correctly, which is why our moral beliefs can be true or false. Secondly, Smith (164–174) also argued that your own moral judgments are not merely about what the idealized, fully rational version of you would want you to do in your actual circumstances, but rather they are about what everyone's fully rational versions would advise their actual, less than fully rational versions to do in your circumstances. This stipulation is required, according to Smith, so that we have a common subject-matter when we debate what we have reasons to do.

Smith (§5.10) also claimed that the previous analysis can be used to explain the practicality of moral judgments. To see how, we need a more careful formulation of 2 where the *ceteris paribus*-clause is replaced with a rationality-condition. We thus get (61):

> 2* If an agent judges that it is right for her to φ in circumstances C, then either she is motivated to § in C or she is practically irrational.

This claim seems to allow us to explain the practicality of moral judgments within the Humean framework. It is constitutive of rational agents that they have a disposition towards coherent and unified combinations of mental states, and hence insofar as you have conflicting combinations of mental states you are irrational (159). Smith then claimed that the combination of (i) believing that everyone's fully rational versions would want their actual versions to φ in C and (ii) lacking a desire to φ in C yourself is incoherent (1997: 100). After all, you believe that a more informed and better reasoning version of you too wants you to φ in C and yet you do not want to φ yourself. As a consequence, insofar as you are a rational agent, your disposition towards coherence and unity will kick in and produce a desire in you to φ in C. This means that, insofar as you are rational, you will have the motivations that match your moral judgments. Furthermore, this explanation seems fully compatible with the Humean idea that beliefs and desires are distinct existences, and it does not assume any brute, inexplicable necessary connections between beliefs and desires either. The moral problem solved?

I believe that there is a problem at the heart of Smith's solution. This is because, even if the solution does not require any necessary metaphysical connections between beliefs and desires and so in a Humean spirit it takes such states to be distinct existences, the solution still is objectionably anti-Humean because it assumes that there are coherence-relations between beliefs and desires.[3] It is, however, less clear of what such relations between beliefs and desires could consist.

According to Smith, beliefs have the mind-to-word direction of fit – they 'purport to represent the way the world is (7).' This representational aspect of beliefs enables us to explain why two beliefs are either consistent with each other or contradict one another (see Fullhart and Martinez (2024)). Roughly put, two beliefs are consistent if there is a possible world in which both are true and inconsistent when there is no such a world (that is, when the truth of one of the beliefs excludes the truth of the other in the same world). We can similarly explain when two desires, as states that represent how the world is to be, are either coherent with one another or conflicting with each other. Again, roughly, two desires are coherent with one another if there is a world in which they are both satisfied, and conflicting when there is no such world because satisfying one of the desires rules out satisfying the other.[4]

Smith's (1997: 100) solution to the moral problem, however, requires that, in addition to these intra-belief and intra-desire relations of coherence and incoherence, it would also make sense to talk about whether a given belief either coheres or conflicts with a given desire. That is, his view requires that we can meaningfully say whether a given representation of how the world is coheres or conflicts with a given representation of how the world is desired to be. This is required because Smith's solution to the moral problem is based on the idea that your belief that our fully rational selves would want our actual versions to φ in C would cohere more with your desire to φ in C than with your lack of a desire to φ in C (or desire not to φ in C). Now, I agree with Smith that, *intuitively*, it seems like here it would be more coherent, given the content of your belief, to have the former desire rather than the latter one. But, nowhere in *The Moral Problem* do we get an explanation of why that former combination of a belief and a desire would be more coherent than the latter combination, and we never get an account of what coherence between beliefs and desires would consist of more generally.[5] We never get

3 Hume rejects necessary connections between distinct existences in Hume (T: 1.3.14.35) and is famously sceptical about coherence-relations between beliefs and desires too (T: 2.3.3.5).

4 Geoffray Sayre-McCord (1997: 75–76) objected to the idea that there are normatively significant coherence and unity relations between desires. For a response, see Smith (1997: §4).

5 For a different way to develop this same problem, see Sayre-McCord (1997: 74). For Smith's attempt to respond to Sayre-McCord's concern 'rather swiftly', see Smith (1997: 101). Smith's response assumes that the objection is based on the concern that there cannot be normatively significant coherence and unity relations between different

an answer to the question of under what conditions would a given belief and desire pair be either consistent or inconsistent with one another. This makes me believe that Smith's view replaces one traditional metaethical mystery (necessary connections between beliefs and desires as distinct existences) with a new metaethical mystery, the required coherence and incoherence relations between beliefs and desires.

This gap in Smith's solution can be used to make sense of several more recent developments in metaethics. Firstly, it allows the non-naturalist realists to attempt to give a similar explanation of the practicality of moral judgments.[6] A non-naturalist can claim that moral judgments are beliefs about *sui generis* moral properties. She can then claim that the content of the belief that φing in C has the non-natural property of rightness is such that desiring to φ in C coheres better with this belief than lacking that desire. Like Smith, the non-naturalist can then argue that the coherence-relation here is *intuitive* and cannot be explained in any other terms. This means that Smith's solution to the moral problem does not seem like an improvement to the solutions that are available for the non-naturalist realists.

There are two other reactions one might have to the previous objection to Smith's view, which would both try to rely on the traditional coherence-relations to explain the practicality of moral judgments. Some expressivists would argue that the moral judgment that it is right to φ in C (i.e., the 'belief' that our fully rational selves would want our actual versions to φ in C) is in fact at least in part a desire-like attitude. It could, for example, be a combination of a desire to be a certain kind of an improved version of oneself and a belief that such an improved version of oneself would want us to φ in C.[7] As a

---

desires. Smith (ibid.) then suggests that, if there are such coherence relations between desires, they will explain the relevant coherence relation between the relevant belief and desire, but this just does not seem to be the case given the different directions of fit of such states.

That the relevant coherence and incoherence relations between beliefs and desires are problematic is furthermore supported by the fact that the standard general tests for coherence and incoherence relations between mental states seem to fail to recognise them. For example, let's assume that a belief is satisfied when true and a desire when the world comes to be so that it fits the way the desire specifies. In this case, we can think that two mental states are coherent when there is a possible world in which both states are satisfied simultaneously (see Fullhart and Martinez (2024: 317)). Consider then a case in which an agent believes that our fully rational versions would want us to φ in C. In this case, there are possible worlds in which this belief is true (and thus in which it really is the case that our fully rational versions want us to φ in C) (i) some in which our desire to φ in C is satisfied (given that we φ in C) and (ii) some in which our desire not to φ in C is satisfied (given that we do not φ in C). This means that, on this test, both the desire to φ in C and the desire not to φ in C are equally coherent with the belief that our fully rational versions would want us to φ in C because the satisfaction of neither of those desires is ruled out by the truth of that belief. This means that, according to this test, the belief in question cannot cohere any more with either one of these desires.

6 See, e.g., Scanlon (2014: 65–66), and for a critical discussion Dreier (2015: 166).

7 See, e.g., Ridge (2014: ch. 4).

consequence, the hybrid expressivists who hold the previous view would be in a position to offer *intra-desire* explanations of why when you judge that it is right to φ in C, having a desire to φ in C would be more coherent. This, however, is just a more complicated way of saying that expressivists can explain the practicality aspect of our moral judgments (though Smith would presumably question whether they can explain the objectivity aspect too).

By contrast, some cognitivists would argue that, when you judge that it is right to φ in C, the corresponding desire to φ in C is in fact deep down some kind of a belief with the mind-to-world direction of fit, perhaps the belief that you have reason to φ in C.[8] As a consequence, these cognitivists would be in a position to offer an intra-belief explanation of why having a desire to φ in C is more coherent when you judge that φing in C is the right thing to do, and thus why, as a rational agent, you would have that desire if you made the moral judgment in question. Yet, the problem with this view response is that it seems to reject the Humean picture of human psychology, the idea that beliefs and desires are distinct existences.

Overall, there is then a worry that Smith has failed to make the seemingly inconsistent propositions 1–3 fully consistent with one another, and so it still seems like we need to reject one of those propositions. Furthermore, after 30 years of debating, we seem no closer to a consensus concerning which proposition that should be, and so the metaethical disagreements carry on as before. The rest of this paper suggests that, instead of trying to solve the moral problem, there might be a way of dissolving it. This will, however, require introducing a new big picture to capture the logical space of metaethical views in a different way, which is a task I will turn to next.

3. Metaethics Meets Dispositional Accounts of Belief

Just like Smith grounded his understanding of the metaethical landscape on the foundations of the Humean picture of human psychology, I want to begin by assuming a different big picture of human psychology, namely the dispositional approaches to belief. It is worthwhile to note that this big picture of human psychology and the nature of belief is assumed here merely for the sake of the argument as it enables us both to draw a new map of the metaethical territory (this section) and, with the help of this new map, to dissolve Smith's moral problem (§4). However, my intention is not to defend the dispositional approach. This is because, in the concluding §5, I will suggest that similar arguments to dissolve the moral problem can also be made in the frameworks provided by the other popular views about the nature of belief.

One key difference between the Humean and the dispositionalist big pictures of human psychology is that, whilst the Humean picture assumes that

8 See, e.g., Gregory (2021).

internal representations and internal structures of the mind are fundamental to being in a given belief state, the dispositional picture sees such things as almost irrelevant for being in the state of believing. Rather, according to the dispositional accounts, beliefs consist of (i.e., are nothing but) 'dispositions to act and react in various ways in various circumstances' (Schwitzgebel 2010: 533).[9] This means that, on these views, beliefs can rightly be ascribed to beings solely based on the patterns of their actual and possible behaviours, irrespective of what, if anything, is going on inside of their minds. The events internal to the mind are on this view relevant derivatively and only when they ground and explain the relevant patterns of behaviour.

What then are the relevant dispositions constitutive of a belief that p?[10] This is a question I will attempt to address in more detail below, but according to the traditional forms of dispositionalism, they consist, for example, of dispositions to assent to utterances of p in the right circumstances, to exhibit surprise when it turns out that not p, to assent to q if p implies that q, to depend on p in one's plans and actions, and so on. All of this is very abstract, and so it is helpful to illustrate this view with Gilbert Ryle's famous more concrete example (1949: 135):

> ... to believe that the ice is dangerously thin is to be unhesitant in telling oneself and others that it is thin, in acquiescing in other people's assertions to that effect, in objecting to statements to the contrary, in drawing consequences from the original proposition, and so forth. But it is also to be prone to skate warily, to shudder, to dwell in imagination on possible disasters and to warn other skaters. It is a propensity not only to make certain theoretical moves but also to make certain executive and imaginative moves, as well as to have certain feelings.

Note that, in this quote, Ryle does not equate the relevant belief with a single disposition but rather with a vast number of different kinds of dispositions to do different things. Because of this, some philosophers talk about a 'multi-track' disposition (where the tracts consist of 'abilities, tendencies or pronenesses to do, not things of one unique kind, but things of lots of different things' (Ryle 1948: 118)) or about a single dispositional track that just happens to be very wide (Marcus 1990; Hunter 2011: 238).

We can make two observations about this approach to belief. Firstly, claims about dispositions hold merely *all else being equal*, against a background

---

9 For defences, see, e.g., Ryle (1949), Price (1969), Audi (1994), and Schwitzgebel (2002). An analogy that can be helpful is that, in a similar fashion, it is appealing to think of character-traits as behavioural dispositions (see Schwitzgebel (forthcoming)).

10 The relevant dispositions can be understood in terms of the truth and falsity of conditional statements of the form "If circumstances C hold, then object O will (or is likely to) enter (or remain in) state S' (Schwitzgebel 2002: 250). This allows us to call O entering the state S the manifestation of the disposition, C the manifestation condition, and the event of C obtaining the trigger (ibid.).

of certain defeasible assumptions (Rowbottom 2007; Schwitzgebel 2010: 534). Even if the previous example's skater is disposed to warn others, she may not do so because she either wants to harm others or is too distracted by other things. Yet, despite these deviations from the typical dispositional manifestations, the skater can count as someone who believes that the ice is thin, given the obtaining excusing conditions. There must, of course, be at least some limits on what constitutes an excusing condition, or otherwise we could ascribe any beliefs we wanted to others. It is, however, equally difficult to state exactly what the limits on these excusing conditions would be.

Secondly, a distinction between two different kinds of dispositions in Ryle's example will be crucial for our purposes below. Ryle (1949: 135) distinguishes dispositions 'to make certain theoretical moves' from dispositions to make 'certain executive and imaginative moves, as well as to have certain feelings.' Let us call the former dispositions 'theoretical' and the latter 'practical'. Ryle uses the dispositions to tell oneself that the ice is thin, to object to others if they deny this, and to use the thinness of ice as a premise in deliberation as examples of the former type of dispositions, and the dispositions to skate warily on the ice, to warn others, and to dwell on the worst-case scenarios as examples of the latter type of dispositions. The former dispositions, the theoretical ones, tend to belong to the part in our mental lives that is more sensitive to evidence and argument and also more controlled, self-aware, and thoughtful. By contrast, the latter, the more practical dispositions tend to be more habitual, automatic, uncontrollable, and associative.[11]

Let us then apply this general approach to belief to moral beliefs.[12] Take a subject, call her Sarah, who believes that eating meat is wrong. According to the dispositional approaches to belief, Sarah would thus have a wide range of dispositions to act and react in different circumstances in the ways that are relevant for that belief, where those dispositions would constitute her belief that eating meat is wrong. Firstly, Sarah would have a wide range of theoretical dispositions that would include tendencies to tell herself that eating meat is wrong, to bring up the topic in discussion, to challenge those who deny that eating meat is wrong, to look for positive evidence and arguments for the wrongness of meat-eating, to use the wrongness of meat-eating as a premise in practical reasoning, and so on. As mentioned above, these theoretical dispositions are a part of Sarah's cognitive architecture that is more explicit, controlled, self-aware, and thoughtful. They also seem to

11 For discussions of this contrast, see Gendler (2008a; 2008b), Zimmermann (2007), and Schwitzgebel (2010: 538). In addition to behaviour dispositions, the relevant dispositions also include cognitive and phenomenal dispositions (see Schwitzgebel (2002: 252)).

12 From this point onwards, the phrase 'moral beliefs' should no longer be understood necessarily to mean belief states understood in Humean terms. Rather, below I will use the phrase as a neutral label for being the state, whatever the ultimate nature of that state is, that is required for being able to sincerely assert the corresponding normative sentence (see Schroeder (2008: §5.1)).

correspond generally to Smith's characterisation of the objectivity of moral judgments.

Sarah would, however, also have a wide range of practical dispositions that would include tendencies to choose to eat vegetarian options at restaurants, to cook vegetarian dishes at home, to protest against factory farming, to feel awkward in the presence of meat-eaters, to buy lots of vegetables, and so on. And, as mentioned, many of these dispositions would be more habitual, implicit, automatic, uncontrollable, and associative. These dispositions thus seem to correspond generally to Smith's description of the practicality of moral judgments. Given that Sarah then has both all the theoretical and all the practical dispositions that we would associate with someone who believes that eating meat is wrong, we are therefore inclined to ascribe to her the belief that eating meat is wrong. According to the dispositionalist accounts, her moral belief that eating meat is wrong would then just consist of those dispositions.[13]

However, in addition to the previous kind of cases, there are also ones in which subjects have only some of the relevant dispositions. In metaethical moral psychology, one such case has been discussed extensively, the case of Huckleberry Finn. Nomy Arpaly (2015: 141–142) describes it in the following way:

13 The previous dispositions that constitute Sarah's moral belief could be called the dispositional stereotype, which are the cluster of dispositions we are apt to associate with the moral belief in question (Schwitzgebel 2002: 251). One important practical consequence of this view is that the grounds on which we tend to ascribe beliefs to others, namely the patterns of their outward behaviour, are according to it intimately connected to the constituents of those beliefs, the behaviour dispositions. This also seems to correspond nicely to Smith's (6–7) intuitive observations concerning the practicality of moral judgments in concrete cases. These cases suggest that we tend to ascribe moral principles to others based on their observable behaviour, and so in the absence of the relevant patterns of behaviour we often begin to question what the agent's moral beliefs ultimately are.

This close connection between the grounds for ascribing moral beliefs and the moral beliefs themselves has traditionally led to two well-known objections to dispositionalism that are concerns about the practical implications of the view. Firstly, it is often pointed out that how a person with a given moral belief behaves depends significantly on her other mental states (beliefs, desires, emotions, mood and the like) and so identifying a given moral belief with a simple behaviour disposition is problematic (see Chisholm (1957)). Secondly, it has also been argued that there are cases where the connection between moral beliefs and patterns of behaviour is just too loose for the purposes of the view under consideration. These cases include actors, paralyzed persons, people who live under oppression and censorship, and moral beliefs about very distant matters (see Putnam (1963) and Strawson (1994)). In all these cases, the individuals seem able to have the relevant moral beliefs without any of the relevant patterns of behaviour being present, or vice versa. In response to these challenges, the contemporary dispositionalists have become more liberal and inclusive concerning what types of dispositions are relevant for having a given belief and about in what kind of situations these dispositions need to be manifested (see Schwitzgebel (2002: 259 and 2024: §1.3)). This paper follows this response in the way it takes the relevant dispositions to be very wide multi-track dispositions.

> To make a long story short, Huckleberry Finn, Mark Twain's fictional character, often known as Huck, is a boy portrayed as an ignorant but good person. Huck, who is white, helps Jim, a black slave, escape. As they float together on a raft on the river, Huck experiences what he thinks of as pangs of conscience. He wonders if he is doing something wrong—stealing from Jim's owner, whom he calls Miss Watson. Upon deliberation, Huck is forced to conclude that helping Jim is wrong and resolves to turn him in. However, when a golden opportunity appears to turn Jim in, Huck finds himself psychologically unable to do it.

This case illustrates how sometimes subjects have conflicting sets of theoretical and practical dispositions relevant to belief.[14] Huck seems to have all the theoretical dispositions that seem constitutive of believing that helping Jim is wrong, and yet, at the same time, he also seems to have all the practical dispositions that would seem constitutive of the opposite belief, the belief that helping Jim is not wrong. He is, after all, disposed not to turn Jim in, not to alert anyone of his presence, and so on.

In the framework of dispositionalism about belief, such cases enable us to focus on the question of exactly which dispositions are constitutive of a given moral belief. One significant consequence of this question is that different answers to it seem to provide a new map to the logical space of different metaethical views. The traditional well-known metaethical views can be located from this map, but it also creates space for new views.

The first response to the previous question is the so-called *pro-judgment view* (Zimmerman 2007; Gendler 2008a and 2008b). On this view, only the theoretical dispositions constitute a belief, mainly because those dispositions, just as beliefs, are thought to belong to the part of our cognitive lives that is rational, thoughtful, and sensitive to evidence and argument (see Schwitzgebel (2010: 538)). In the previous case, this view entails that Huck's theoretical dispositions would constitute his belief that helping Jim is wrong, whereas his practical dispositions would be both irrelevant for having that belief and fail to constitute a belief that helping Jim is not wrong. In terms of traditional metaethical views, this view would most naturally correspond to different forms of externalist cognitivism to which Smith (68–76) has always objected.[15]

The second response is the so-called *anti-judgment view* (Hunter 2011). According to it, only the practical dispositions constitute beliefs. The main motivation for this view is that, often in the cases in which subjects'

---

14 For Smith's own descriptions of such cases, see, e.g., (67) and Smith and Kennett (1994 and 1996). Schwitzgebel (2010: 532–533) likewise describes the cases of Juliet the implicit racist, Kaipeng the trembling Stoic and Ben the forgetful driver in which the agents' theoretical and practical dispositions come apart.

15 For objections to the pro-judgment view more generally, see Schwitzgebel (2010: 538–541).

theoretical and practical dispositions conflict, intuitively we tend to ascribe beliefs based on the practical dispositions because they seem to better match what we take the agents cognitive stance to be (Schwitzgebel 2010: 541). In the previous case, this view would entail that Huck's practical dispositions to help Jim constitute his belief that doing so is not wrong, whereas his theoretical dispositions are neither relevant for that belief nor constitute a belief that helping Jim is wrong. In terms of traditional metaethics, this view would most naturally correspond to different forms of internalist non-cognitivism and expressivism to which Smith (2001 and 2002) has always objected too.[16]

The third response we could call the *belt and suspenders view*. On this view, all the theoretical and practical dispositions together constitute a given belief, but only when they all perfectly align with each other. This view would entail that, in the case above, Sarah's coinciding theoretical and practical dispositions do successfully constitute her belief that eating meat is wrong. By contrast, because Huck's theoretical and practical dispositions are pulling in the opposite directions, Huck would on this view neither believe that helping Jim is wrong nor that it is not wrong. In terms of traditional metaethics, this view would most naturally correspond to different forms of internalist cognitivism.[17]

In the framework of dispositionalism, we can also, at this point, construct new, previously unexplored metaethical positions. For example, according to the *shifting view*, subjects like Huck are shifting between having different beliefs (Rowbottom 2007). On this view, when Huck is in reflective contexts, he has a high degree of belief that helping Jim is wrong, and yet, when he moves to a non-reflective context where action is called for, he loses this belief and perhaps even slides to believing that helping Jim is not wrong. This means that, on this view, the dispositions that are manifested in a context constitute the belief the subject holds in it.[18]

Another answer would be provided by the *contradictory view* (Gertler 2008). On this view, both theoretical and practical dispositions constitute separate beliefs of their own. According to this view, given his conflicting theoretical and practical dispositions, Huck would believe both that helping Jim is wrong and that it is not wrong. This is because both his sincere avowal

16 For objections to the anti-judgment view more generally, see Schwitzgebel (2010: 541–543).

17 With respect to the theories of this type, Smith (118–125) objects to the non-Humean versions that seem based on 'besires' in an objectionable way. His own view too, however, is, as explained above, a version of cognitivist internalism that just promises to be compatible with Humean moral psychology. The view described in the next section will be closest to this view, although (unlike the high threshold view) it will not require that subjects who have a given moral belief have all the theoretical and practical dispositions relevant to it.

18 Schwitzgebel (2010: 543–544) objects to this view on the grounds that it makes us unable to describe an agent's overall, general attitude in the cases of the conflicted agents. For example, it leaves it open what Huck really believes when he is neither deliberating nor in a position to help Jim.

that helping Jim is wrong and his spontaneous reaction not to help him are sufficient on their own to underwrite belief.[19] Interestingly, the shifting view and the contradictory view do not have natural counterparts in the traditional metaethical debates.

This suggests that (i) dispositionalism, (ii) the cases of conflicted agents, and (iii) the question of which dispositions in them constitute beliefs together provide us with a new map of the metaethical landscape from which both the traditional metaethical views and new alternatives can be located. This map is in several ways different from the one provided by Smith's moral problem. The real question then is whether the new map is any better than the old map. Is it, for example, in any way more useful or illuminative?

The problem with the old map is that, assuming that Smith's own solution to the moral problem suffers from the issue explored in §2, we still seem to face the intractable question of whether we should reject (i) the objectivity moral judgments, (ii) their practicality, or (iii) the Humean picture of human psychology. And, 30 years on, it seems like we are no closer to a generally accepted answer to this question.

By contrast, in terms of the new map, a solution to the moral problem would consist of a definitive answer to the new fundamental metaethical question of which dispositions (theoretical, practical, both together or separately on their own, or some other alternative) constitute a given moral belief. As I suggested above, this question too can be used to distinguish between different metaethical positions, and so presumably the defenders of the traditional metaethical views would be inclined to defend different answers to this question as well.

There are, however, two reasons to believe that, in this form, this new metaethical problem will be just as intractable as Smith's moral problem. Firstly, in the more general debates about beliefs and also in the debates about implicit biases more specifically, all the previous answers to the question of which dispositions constitute a belief continue to be equally controversial. Thus, at least sociologically speaking, the debates about which dispositions constitute beliefs seem just as intractable as the old metaethical debates.

Secondly, in the traditional metaethical debates between the so-called motivational internalists and externalists, the cases in which agents' theoretical dispositions and practical dispositions come apart have been discussed intensively for quite a while now.[20] In these debates, the defenders of the different positions have different intuitions about the relevant cases, and they also give very different descriptions of them to match their theories. Because of this, there is little hope that we could come to agree on the question of which dispositions constitute moral beliefs just by consulting

19 Schwitzgebel (2010: 544) suggests that this view does not add anything of value besides confusion.

20 For an outline of these debates and references, see Björklund et al (2012).

our intuitions about the relevant test cases. This, furthermore, means that it is unlikely that we could use the new map to converge on the correct metaethical view, whatever that may be. Because of this, I am sceptical about whether the new map is any more useful than Smith's map with respect *to solving* the most fundamental metaethical problems. The next section will, however, suggest that, just maybe, the new metaethical map could turn out to be more helpful in a different way. Perhaps instead of helping us to solve Smith's moral problem, the new map will help us to dissolve many of the most fundamental metaethical problems by helping us to see why those questions cannot be answered in the first place.

4. Dissolving the Moral Problem

Dissolving the moral problem in the framework of dispositionalism about belief requires making two theoretical moves. We first need to understand the dispositions that constitute different beliefs in a much more fine-grained way, and we then need to take 'believe' to be a vague predicate that admits of so-called 'in-between' cases. This section will first outline these two moves. It will then explain how they will lead, in two ways, to the dissolution of the moral problem.

The previous section focused on two separate sets of dispositions relevant to belief, the theoretical dispositions and the practical dispositions. This discussion used examples of both kinds of dispositions, but, in a very coarse-grained fashion, it also gave the impression that the theoretical dispositions and the practical dispositions always exist separately as unified and complete sets.

We should, however, think of the relevant dispositions in a much more fine-grained way. Firstly, we should really talk about thousands of narrower, more local dispositions. For example, Sarah might have separate dispositions to make different choices in different restaurants, shops, and kitchens, different dispositions to react to people eating meat in different situations, different dispositions to make different arguments in different debates, and so on.[21] Secondly, we should not think that these dispositions are either fully theoretical or fully practical, but rather we should think that they are on a spectrum of theoreticality and practicality. This is because these dispositions can be more or less controlled/uncontrollable, self-aware/automatic, deliberative/associative, explicit/implicit, thoughtful/reactive, and action-/argumentation-orientated.

Thirdly, and most importantly, we should accept that a subject can have these dispositions in any combination whatsoever and not just as full sets of theoretical and practical dispositions. As Schwitzgebel (2010: 534) puts it, '[t]here must, indeed, be something like a continuum between full

21 For many beliefs, the relevant dispositions would include a vast number of dispositions to take and refuse different bets in different situations (Ramsey 1931).

possession of all the relevant dispositions and possession of none of them – with a multidimensional spectrum of cases between the two extremes.' By this, Schwitzgebel means that if we take a subject who has a certain percentage of the dispositions (say 64% of them) relevant to having a given, these dispositions could be any of the hundreds of more or less theoretical and practical dispositions that are relevant for having the belief in question. These dispositions could be almost wholly from the more theoretical end of the spectrum, almost wholly from the more practical end of the spectrum, from the middle of the spectrum, or any other mix of the theoretical and practical dispositions. This follows from the Humean dictum that there are no necessary connections between distinct existences (which in this case are all the different dispositions that are relevant to having the belief in question) (Hume T: 1.3.6.1).

In order to make the second theoretical move, we can then focus on the predicate 'believe'. If we adopt the previous picture of a continuum of dispositions with a multidimensional spectrum of cases between the extremes, it will be natural to think that it is a 'vague predicate that admits of in-between cases' (Schwitzgebel 2010: 533). This entails that there is no simple answer to the question of which dispositions constitute a given belief. There will not be any specific dispositions that will be individually necessary and jointly sufficient for having that belief. Rather, the following picture emerges. If a subject such as Sarah has most of the dispositions relevant to believing that eating meat is wrong, she will definitely believe that eating meat is wrong. In other words, in that case the mix of both practical and theoretical dispositions Sarah has constitute her belief that eating meat is wrong. At the opposite end of the spectrum, if a subject has only a few of the relevant dispositions, she will definitely lack the belief that eating meat is wrong as she would not have enough of the relevant dispositions to constitute that very belief.

There is, however, a broad range of cases in-between these two ends of the spectrum, where we can in principle fully articulate in detail the subject's dispositional structure. In these cases, we could in principle list comprehensively the combination of the relevant theoretical and practical dispositions the subject has (and likewise the dispositions she lacks). Here it is, however, natural to think that at this broad middle zone of the multidimensional spectrum there is no sharp threshold at any point that would partition the spectrum into cases of believing and not believing (Schwitzgebel 2010: 535). It seems much more plausible that there is a wide zone of the spectrum in the middle that contains the cases that could be called the 'in-between' cases in which it is indeterminate whether the subject has the given belief. These are cases of vagueness in which, even if we could in principle know the subject's dispositional structure in fine detail, even that knowledge would still require us to refrain from either ascribing or denying the belief in question. In these cases where the relevant dispositions are only partially possessed, the talk about beliefs begins to break down as 'the simplifications

and assumptions inherent in it aren't entirely met' (Schwitzgebel 2010: 535). Here, there just is no fact of the matter – it's neither true nor false that the subject in question has the relevant belief.

In this section, I have thus outlined two moves that lead to a new dispositional picture of moral beliefs. This picture is based on first thinking of the dispositions that are relevant to belief in a more fine-grained way. It thus recognises that the relevant fine-grained dispositions can be more or less theoretical and also that they are distinct existences that can be had in a multitude of different combinations. We thus get a spectrum where at one end are the cases of having most of the relevant dispositions (the cases of belief), and at the other end the cases of having only few of the dispositions (the cases of disbelief). And, in the middle, we have a broad range of in-between cases, i.e., the indeterminate cases. I then want to suggest that this picture enables us to dissolve the moral problem in two ways.

Firstly, according to Smith's moral problem, if we assume the Humean picture of human psychology (and reject Smith's own solution), we face the choice between externalist cognitivism (which requires giving up the practicality of moral judgment – thesis 2 in §1 above) and internalist non-cognitivism (which requires giving up the objectivity of moral judgment – thesis 1 in §1 above). This forced choice has led many metaethicists to consider cases like Huck Finn above in which agents have all the theoretical dispositions relevant to a given moral judgment and none of the corresponding practical dispositions. This has furthermore led to extensive discussions of cases concerning amoralists, psychopaths, depressed individuals, and evil people.[22] The hope has been that, by coming to a view about whether these agents have made genuine moral judgments, we would be able to decide between externalist cognitivism and internalist non-cognitivism.

The dispositionalist picture of belief can, however, explain why an agreement cannot be reached about the previous kind of cases and hence also why the debate between the externalist cognitivists and the internalist non-cognitivists will never be able to come to a conclusion. Recall that, according to that picture, if a subject has most of the (theoretical and practical) dispositions relevant to a given moral belief those dispositions will successfully constitute the belief in question; if she only has few of them she fails to have the belief; and in between these ends of the spectrum there is a wide middle zone in which the subject is in between believing and disbelieving – where it is indeterminate whether the subject has the relevant belief. If this is the case, it seems likely that the test cases that have traditionally been used in the cognitivism versus non-cognitivism debate will fall into this middle zone. After all, in these cases, the subjects have many of the theoretical dispositions and hardly any of the practical ones (or vice versa), and so they neither have a majority of all the relevant dispositions nor lack the majority of them. This

22 For an overview, see Björklund et al (2012).

means that the cases of Huckleberry Finn and the familiar cases of amoralists, psychopaths, depressives, and evil individuals are likely to be indeterminate cases in which the subjects are in-between having the relevant moral beliefs and not having them. In these cases, given the subjects' dispositions, there just is no fact of the matter. This would entail that the debate between the externalist cognitivists and internalist non-cognitivists is in principle unsolvable, and so the question these views are trying to answer seems to be go away – it seems to dissolve.

There is also, however, a second, deeper reason for why the dispositionalist view of moral belief dissolves the moral problem and, with it, many of the traditional metaethical questions. To see this, we need to consider the metaphysical status of beliefs in Smith's framework and compare it to the dispositional picture. Here it is helpful to consider an analogy.

If we assume the Humean picture of human psychology adopted by Smith, we can think of beliefs and desires as particular mental entities that have their own unitary existence (119).[23] To illustrate this, we can use the analogy of basic physical particles such as electrons and protons and their qualities such as charge. A single particle can have either a positive charge or a negative charge (or neither), but it is impossible for it to have both charges at the same time. Likewise, it can be thought that a single mental state can either have the mind-to-world direction of fit of beliefs or the world-to-mind direction of fit of desires, but not both directions of fit at the same time. As Smith (119) puts it, at least modally it must be possible to be in any belief-state whatsoever without at the same time being at any particular desire-like state. This is why Smith thought that there cannot be 'besires', states with both directions of fit at the same time. Within this framework, it then becomes natural to ask whether a moral judgment is a single, unitary belief-state or a single, unitary desire-state. It is this question that led to the debate between the externalist cognitivists and internalist non-cognitivists and the moral problem as Smith conceived it.

Now, since *The Moral Problem* was published, a third alternative has been explored. According to the hybrid views, we should not think of moral beliefs with the analogy of the basic particles, but rather with the analogy of whole atoms that consist of those particles in some combination. Each of the individual mental states that together constitute a given moral belief is either a belief-state with the mind-to-word direction of fit or a desire-state with the world-to-mind direction of fit, and yet the moral belief itself constituted by those states is neither a distinct unitary belief-state nor a distinct unitary desire-like state but rather a combination of such states. Thus, for example, according to Michael Ridge's (2008: 55) expressivist version of this type of a hybrid view,

23 Hume seems to adopt this type of an atomistic picture in his account of ideas and impressions (T: 1.1.1). However, Blackburn (2008: 19–20) interestingly provides a causal, functionalist interpretation of the difference between impressions and ideas, which Hume (T: 1.1.1) took to be less vivid copies of the former.

a moral belief consists of approval/disapproval of actions (a desire-like state) insofar as they have a certain property and a belief that the actions the moral belief is about have that property.[24] The defenders of such hybrid views believe that one advantage of such ecumenical views is that they promise to explain both the objectivity and the practicality of moral judgments.

The dispositional picture of moral beliefs is, however, a more radical departure from the atomism assumed by Smith's Humean picture of human psychology. According to the new picture, a subject's moral belief is not metaphysically a unitary state but rather it consists of a multitude of more or less theoretical and practical dispositions relevant to having the belief. These constituents of the belief are individually neither beliefs nor desires because they have neither the world-to-mind nor the mind-to-word direction of fit. Rather, they are merely individual dispositions to react in different circumstances in different ways, where some of these dispositions are more controlled, self-aware, thoughtful, and more connected to deliberation and arguments, and others more habitual, automatic, uncontrollable, associative, and manifesting themselves as concrete actions.

This means that, in this framework, the question of whether a given moral belief is fundamentally a unitary belief-state like the cognitivists think, a unitary desire-like state like the non-cognitivists think, or a combination of unitary belief– and desire-states as the hybrid theorists think dissolves. None of these options accurately capture the constituents of moral beliefs given that such beliefs consist of rich combinations of different dispositions (that can still ground the objectivity and practicality aspects of moral judgments Smith was so keen preserve).[25] In this way too, the dispositionalist picture of the human psychology outlined above dissolves the traditional questions of the nature of moral judgments, which led to Smith's formulation of the moral problem and to the intractable debate between the cognitivists and non-cognitivists.

---

24 According to Ridge, this is an expressivist view because the truth of the factual belief does not determine whether the moral belief in question is true. For a cognitivist version of the hybrid views according to which that is the case, see, e.g., Copp (2001).

25 Interestingly a given sufficient mix of the relevant dispositions that constitute a certain moral belief in a given case can contain either a majority of more theoretical or more practical dispositions. This means that, of some cases, traditional externalist cognitivist can be closer to the true picture, whereas of other cases traditional internalist non-cognitivism can be closer to the truth.

It could be objected at this point that the proposed view cannot make sense of the objectivity aspect of moral judgments. It could be argued that dispositions are not the kind of things that can represent objective facts and be either true or false and so, if we thought moral beliefs consisted of dispositions, they could not represent objective facts or be true or false either (see Quilty-Dunn and Mandelbaum (2018: §3.3)). The dispositionalists do not, however, deny that beliefs have propositional truth-evaluable content. Rather, they merely are giving an account of what having a belief with a certain content consists. This means that, according to dispositionalism, the content of the belief in question can still be either true or false in a robust way, which explains also why the relevant more theoretical dispositions are in part constitute of having the belief in question.

5. Conclusion and Extending the New Picture to Other Accounts of Belief

This paper has argued for the following claims. Firstly, even if the three jointly inconsistent claims of Smith's moral problem have provided a hugely influential map of the metaethical landscape, the problem itself seems just as intractable today as it did 30 years ago. Secondly, Smith's own solution to the problem is problematic because it relies on positing brute relations of coherence and incoherence between beliefs and desires. Thirdly, dispositionalism about belief and the question of which dispositions are constitutive of moral beliefs can be used to construct a new map of the metaethical logical space. As we saw, we can locate the familiar traditional metaethical views on this map, but it also creates space for new, previously unexplored positions.

Finally, the previous section suggested more radically that the new way of seeing the metaethical landscape allows us to dissolve the moral problem. To this end, we must first think of the different dispositions relevant to moral beliefs in a more fine-grained way, and we must take 'believe' to be a vague predicate that allows in-between cases in which a subject neither holds a certain moral belief nor determinately lacks it. If this broad picture is right, then, firstly, it is likely that the traditional cases that have been used in the externalist cognitivism versus internalist non-cognitivism debates belong to the category of the in-between cases. In these cases, there is thus no fact of the matter whether the agent has the relevant moral belief or not. And, more fundamentally, on this view, as moral beliefs are taken to consist of a vast number of different dispositions, the question of whether a moral belief is a unitary belief-state, a unitary desire-like state, or some combination of such states just falls away as a question to which no answer can be given.

There is, however, an objection that many would want to make at this point.[26] It could be argued that the previous way of dissolving metaethical questions assumes the dispositionalist view of belief, which admittedly is controversial. I want to conclude by suggesting that this is not quite right. I have relied on dispositionalism merely for the sake of simplicity. Similar attempts to dissolve the moral problem could also be made in the frameworks provided by the three most popular theories of belief: functionalism, representationalism, and interpretationalism as they too create room for cases of in-between believing (see Schwitzgebel (2010: 535–536)).

According to functionalists, believing that P consists of being in a state that occupies (or is apt to occupy) a certain causal-functional role.[27] Smith (113) himself seems to accept this type of a view of beliefs and desires. According to him, the constitutive functional role of belief-states is that the belief that P tends

26 For several other objections, see footnotes 13 and 25 above.

27 See, e.g., Putnam (1967), Armstrong (1968), and Lewis (1972).

to go out of existence in the presence of perception with the content that not P (115). Yet, if we want to describe the functional role of beliefs more generally, the belief that P also tends to be 'brought about by perceiving or hearing about or inferring that P', it tends to 'lead to avowals of P', it tends to 'promote action A if it is discovered that A will achieve a desired goal if P is true', and it tends to be combined with the belief *if P then Q* to conclude that P (Schwitzgebel 2010: 535–536). In this framework, it is natural to think that subjects can be in states that only partially match the previous functional role of beliefs. If we then again take 'believe' to be a vague predicate, in the relevant cases of partial match the subject can be understood to be in-between believing and not believing. This is natural especially if, following Smith (113), we take the functional roles of beliefs to consist of causal dispositions.

According to the so-called representationalists by contrast, believing that P consists of possessing, 'in a belief-like way, an internal representational token (perhaps a sentence in the language of thought) with the content P' (Schwitzgebel 2010: 536). Yet, in a more general sense of the term, the representationalists too are arguably committed to a form of functionalism. This is because most representationalists think that what makes a given representational state the belief it is consists, not only of the further cognitive relationships the state is apt or likely to enter (as per standard functionalism), but also of the facts about how that particular state came about and of the evolutionary or developmental history of that kind of states in the organism or the species (ibid.).[28] As a consequence, in this framework too, it seems natural to think that subjects can be in states that only partially fill the relevant functional roles in the broader sense. In these cases, the subjects would again be in-between believing and not believing – in-between possessing the relevant internal representational token and not doing so.[29]

---

28 See also Fodor (1968; 1987), Millikan (1984), Lycan (1986), Dretske (1988), and Nichols and Stich (2003). Some representationalists defend psychofunctional versions of representationalism (Quilty-Dunn and Mandelbaum 2018). According to these views beliefs are relations to structured mental representations where being in such a relation is determined by the holding of certain generalisations regarding belief acquisition, storage, and change. The defenders of these views have two main objections to the dispositionalist picture discussed in this paper. Firstly, they argue that the cases of in-between beliefs can be understood with the notion of fragmented beliefs where distinct fragments of conflicting beliefs can be stored simultaneously in distinct architectural locations in the same brain (ibid: 2358–2359). Secondly, they also argue that mental representations that are physically realized in the brain are required to play a role in explaining several features of beliefs such as their ability to cause behaviour and enable agents to sort things, their opacity and truth-evaluability, their relations to other propositional attitudes, belief change, and so on (ibid.: §3). Even the defenders of such views, however, grant that we 'should expect ordinary yes-or-no belief ascription to fail in a wide range of cases' (ibid.: 2360), which is sufficient for the purposes of dissolving the moral problem in the way outlined above. For the dispositionalists objections to this type of representationalism generally, see Schwitzgebel (forthcoming).

29 The representationalists often use metaphors such as the relevant representations being in 'belief boxes, 'memory stores' and 'file folders' that suggests a binary architecture, but

It is important to note that, even if this second framework too would allow us to dissolve many traditional questions in metaethical moral psychology, it would still also enable us to ask many of the traditional metaethical questions concerning the relevant representations and their content. For example, we could still have a debate about whether that content is provided by some *sui generis* non-natural facts or by some ordinary natural facts either realistically or relativistically construed. In fact, Smith himself could still argue that his analysis of the content of moral claims (see §2 above) must be correct, because only it can explain why both theoretical and practical dispositions are constitutive of the functional roles of our moral beliefs. In other words, in the representationalist framework, Smith's account of the content of our moral beliefs based on what our fully rational selves would want us to do could be used to explain why both (i) the more controlled, self-aware and thoughtful dispositions related to deliberation and rational argument and (ii) the more habitual, automatic, incontrollable, and associative dispositions more directly related to action are relevant to having a given moral belief.

Finally, according to the interpretationists, if a subject believes that P, this belief consists of exhibiting certain patterns of behaviour that, based on the interpretative tools of folk psychology and the principle of charity, can be made sense of by attributing the subject in question the belief that P (Schwitzgebel 2010: 536).[30] Yet, this framework too leaves room for the relevant cases of in-between beliefs. This is, for example, already because a given actual pattern of behaviour can more or less match the previous type of patterns, and also because the attribution of the relevant belief can more or less make sense of the subject's behaviour overall.

Schwitzgebel (2010: 536) thus more generally concludes that:

> '[o]n all of the leading contemporary approaches to belief, it's natural to suppose that there will be a wide array of in-between cases where the dispositional or functional or functional-historical role is only partially filled, the relevant patterns of behaviour, response and cognition only partly possessed.

This suggests that the arguments of §3–§4 can be adapted to these approaches to belief as well. In these frameworks too, it seems likely that the test cases used in the externalist cognitivism versus internalist non-cognitivism debate will fall under the category of the indeterminate in-between cases in which the subject neither holds nor lacks the relevant belief. And, in all these frameworks, it seems that moral beliefs do not fundamentally consist of unitary belief-states, unitary desire-states, or some combination of the two but rather of having some combination of dispositions, some more theoretical and some more practical than others, or so I have suggested.

---

this is more of a feature of the metaphors than a structural consequence of the theories themselves (Schwitzgebel 2010: 536).

30 See, e.g., Davidson (1984) and Dennett (1987).

References

Armstrong, David 1968, *A Materialistic Theory of the Mind*. London: Routledge.

Arpaly, Nomy 2015. 'Huckleberry Finn Revisited: Inverse Akrasia and Moral Ignorance', in R. Clarke, M. McKenna, and A. M. Smith, eds., *The Nature of Moral Responsibility: New Essays.* Oxford: Oxford University Press, 141–156.

Audi, Robert 1994. 'Dispositional Beliefs and Dispositions to Believe', *Noûs* 28: 410–434.

Blackburn, Simon 2008. *How to Read Hume*. London: Granta.

Björklund, Fredrik, Gunnar Björnsson, John Eriksson, Ragnar Francén Olinder, and Caj Strandberg 2012, 'Recent Work on Motivational Internalism', *Analysis* 72: 124–137.

Chisholm, Roderick 1957. *Perceiving*. Ithaca, NY: Cornell University Press.

Copp, David 2001. 'Realist Expressivism – A Neglected Option for Moral Realism', *Social Philosophy and Policy* 18: 1–43.

Davidson, Donald 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Dennett, Daniel 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dreier, James 2015. 'Another World: The Metaethics and Metametaethics of Reasons Fundamentalism', in R. N. Johnson and M. Smith, eds., *Passions & Projections – Themes from the Philosophy of Simon Blackburn*. Oxford: Oxford University Press, 155–171.

Dretske, Fred 1988. *Explaining Behaviour*. Cambridge, MA: MIT Press.

Fodor, Jerry 1968. *Psychological Explanation*. New York: Random House.

Fodor, Jerry 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Fullhart, Samuel and Camilo Martinez 2024. 'Coherence as Joint Satisfiability'. *Australian Journal of Philosophy* 102: 312–332.

Gendler, Tamar Szabó 2008a. 'Alief and Belief', *Journal of Philosophy* 105: 634–663.

Gendler, Tamar Szabó 2008b. 'Alief in Action (and Reaction)'. *Mind & Language* 23: 552–585.

Gertler, Brie 2008. 'Self-Knowledge and the Transparency of Belief', in A. Hatzimoysis, ed., *Self-Knowledge*. Oxford: Oxford University Press, 125–145.

Gregory, Alex 2021. *Desire as Belief – a Study of Desire, Motivation & Rationality.* Oxford: Oxford University Press.

Hunter, David 2011. 'Alienated Belief', *Dialectica* 65: 221–240.

Hume, David 1978 [1939–40]. *A Treatise of Human Nature*, 2nd edn., eds. L.A. Selby-Bigge & P.H. Niddich. Oxford: Clarendon Press.

Lewis, David 1972. 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy* 50: 249–258.

Lycan, William 1986. 'Tacit Belief', in R.J. Bogdan, ed., *Belief*. Oxford: Oxford University Press, 61–82.

Marcus, Ruth 1990. 'Some Revisionary Proposals about Belief and Believing', *Philosophy and Phenomenological Research* 50: 132–153.

Millikan, Ruth 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Nichols, Shaun and Stich, Stephen 2003. *Mindreading*. Oxford: Oxford University Press.

Price, H.H. 1969. *Belief*. London: George Allen & Unwin.

Putnam, Hilary 1963. 'Brains and Behavior', in R. Butler, ed., *Analytical Philosophy*. Oxford: Basil Blackwell, 1–19.

Putnam, Hilary 1967. 'Psychological Predicates', in W. H. Capitan and D. D. Merrill, eds., *Art, Mind and Religion*. Pittsburgh: University of Pittsburgh Press, 37–48.

Quilty-Dunn, Jake & Eric Mandelbaum 2018. 'Against Dispositionalism: Belief in Cognitive Science', *Philosophical Studies* 175: 2353–2372.

Ramsey, Frank 1931. 'Truth and Probability', in R.B. Braithwaite, ed., *The Foundations of Mathematics and other Logical Essays*. London: Kegan, Paul, Trench, Trubner & Co., 156–198.

Ridge, Michael 2014. *Impassioned Belief*. Oxford: Oxford University Press.

Rowbottom, Darrell P. 2007. '"In-Between Believing" and Degrees of Belief', *Teorema* 26: 131–137.

Ryle, Gilbert 1948. *The Concept of Mind*. New York: Barnes & Noble.

Sayre-McCord, Geoffray 1997. 'The Metaethical Problem', *Ethics* 108: 55–83.

Scanlon, T.M. 2014. *Being Realistic about Reasons*. Oxford: Oxford University Press.

Schroeder, Mark 2008. 'Expression for Expressivists', *Philosophy and Phenomenological Research* 76: 86–116.

Schwitzgebel, Eric 2002. 'A Phenomenal, Dispositional Account of Belief', *Noûs* 36: 249–275.

Schwitzgebel, Eric 2010. 'Acting Contrary to Our Professed Beliefs or the Gulf between Occurrent Judgment and Dispositional Belief', *Pacific Philosophical Quarterly* 91: 531–553.

Schwitzgebel, Eric 2024. 'Belief', *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2024/entries/belief/>.

Schtitzgebel, Eric forthcoming. 'Dispositionalism, Yay! Representationalism, Boo!', in Jonathan Jong and Eric Schwitzgebel, eds., *The Nature of Belief*. Oxford: Oxford University Press.

Smith, Michael 1994. *The Moral Problem*. Oxford: Blackwell.

Smith, Michael 1997. 'In Defence of *The Moral Problem*: A Reply to Brink, Copp and Sayre-McCord', *Ethics* 108: 84–119.

Smith, Michael 2001. 'Some Not-Much-Discussed Problems for Non-Cognitivism in Ethics', *Ratio* 14: 93–115.

Smith, Michael 2002. 'Evaluation, Uncertainty and Motivation'. *Ethical Theory and Moral Practice* 5: 305–320.

Smith, Michael and Kennett, Jeanette 1994. 'Philosophy and Commonsense: The Case of Weakness of Will', in M. Michael and J. O'Leary-Hawthorne, eds., *Philosophy in Mind: The Place of Philosophy in the Study of Mind*. Dordrecht: Kluwer, 141–157.

Smith, Michael and Kennett, Jeanette 1996. 'Frog and Toad Lose Control', *Analysis* 56: 63–73.

Strawson, Galen 1994. *Mental Reality*. Cambridge, MA: MIT Press.

Zimmerman, Aaron 2007. 'The Nature of Belief', *Journal of Consciousness Studies* 14: 61–82.

Nathan Howard
University of Toronto
Department of Philosophy
nathan.howard@utoronto.ca

Original Scientific Paper
UDK 17.026.4
17.032:161.224.2
Received: October 02, 2024
Revised: October 31, 2024

Accepted: November 16, 2024

CONVERGENCE AND THE AGENT'S POINT OF VIEW

Abstract

This paper examines the apparent tension in Michael Smith's The Moral Problem between his commitment to convergence in ideal desires and his acceptance of agent-relative reasons, particularly those grounded in first-personal perspectives like the parent-child relationship. While Smith maintains that ideal desires are agent-invariant and converge on what is universally desirable, he also endorses agent-relative reasons that imply agent-centered normative commitments. I argue that resolving this tension requires rethinking convergence. Specifically, I propose extending the first-personal („de se") nature of agent-relative reasons to the objects of convergence, which I term „de se aims." By recognizing these aims as value bearers, we can reconcile agent-relativity with the universality of desirability, preserving Smith's broader metaethical commitments. The proposal avoids the pitfalls of agent-relative value theories and illuminates the role of perspective-dependent aims in systematic justification.

Key words: Convergence · Agent-relative reasons · Agent-neutral value · First-personal perspective

1. Introduction

Two core ideas of *The Moral Problem* seemingly conflict: convergence and agent-relativity. Michael Smith distinguishes between three types of relativity in normative ethics—relativity of normative concepts, agent-relativity of reasons, and circumstantial relativity of reasons. While he rejects the first form, as famously attributed to Bernard Williams, he accepts the latter two. Despite accepting relativity in ethics, Smith also asserts that the ideal desires underlying normative reasons converge. According to this view, which desires are ideal is, seemingly, not an agent-relative matter. This paper asks whether convergence is compatible with agent-relative reasons, particularly those arising from personal relationships like the parent-child bond, which Smith calls first-personal or '*de se*' reasons.

I will argue that Smith can resolve this tension, but only by rethinking the nature of the convergence that he advocates for some ideal desires. In particular, he must extend the source of agent-relativity that he locates in *de*

 DOI: 10.5937/BPA2437145H

se reasons to the desirable objects on which ideal desires converge. I call these objects '*de se* aims' and defend their appeal as points of convergence.

In the first half of the paper, comprising sections one to four, I establish the tension between convergence and agent-relative reasons by explaining both concepts as they figure in *The Moral Problem*. Section two focuses on the role of *de se* thought in both reasons and ideal desires, drawing on a famous argument from G. E. Moore that agent-relative norms, such as egoistic norms, conflict with agent-neutral conceptions of value, assuming consequentialism. The second half of the paper, comprising section five and the conclusion, briefly rejects resolving this tension through agent-relative conceptions of value and argues that we should instead resolve it by relativizing the objects of convergence to agents in the manner that agent-relative reasons are relativized, which does not require relativizing our conception of value.

2. Stage-Setting

Convergence

The Moral Problem defends a conception of normative reasons that overcomes the titular problem, the problem, very roughly, of reconciling the cognitive character of moral judgments with their motivating power. On page 150, after bobbing and weaving through the history of analytic ethics, we arrive at our destination: an analysis of normative reasons. That analysis — platitude-summarizing and non-reductive, of course — has two parts:

1. What it is desirable that we do is what we would desire to do if we were fully rational;
2. What we have normative reason to do is what we would desire that we do if we were fully rational.

From this it follows that what we have reason to do is what it is desirable that we do. This doesn't quite tell us what to do, yet, for we don't yet know what counts as desirable. But light shines from page 152: "facts about what it is desirable for us to do are constituted by the facts about what we would advise ourselves to do if we were perfectly placed to give ourselves advice."

What would I have to be like, were I perfectly placed to advise myself? Elsewhere, Smith writes that "to be fully rational an agent must not be suffering from the effects of any physical or emotional disturbance, she must have no false beliefs, she must have all relevant true beliefs, and she must have a systematically justifiable set of desires, that is, a set of desires that is maximally unified and coherent."[1] These "ideal advisors" have, as Bukowski puts it, "maximally exercised the two capacities that are constitutive of agency, namely, knowledge acquisition and desire realization."[2]

1 Smith (1997: 89)

2 Bukoski (2016: 121)

Speaking strictly of *The Moral Problem*'s impact on the field, perhaps its most controversial idea is that our ideal desires — that is, the desires of these ideal advisors — converge. The notion of some desires being ideal isn't *particularly* controversial. Rather, Smith's claims have provoked such discussion because, according to his conception, this set of ideal desires is also the set that *you* would have, were *you* to maximize your agentive capacities and deliberate correctly, no matter the content of your actual desires. That is, regardless of your starting point, as you come to know more, you will come to desire more ideally. Bernard Williams famously expressed scepticism about convergence in our ideal desires, limiting the process of idealizing desire to the exercise of Humean or, roughly, broad instrumental rationality, thereby limiting the scope of change between actual and ideal desire. But Smith rejects Williams' "scepticism about the scope of reasoned change in our desires (Korsgaard 1986); predicated on denying that, through a process of rational deliberation — through attempting to give a *systematic justification* of our desires, for example — we could ever some to discover reasons that we all share" (1994: 165).

Smith draws this conception of systematic justification from John Rawls and Christine Korsgaard. Williams' contrasting Humean conception of practical reasoning tethers the desires an agent could acquire through sound practical deliberation from their actual desires. This conception of practical deliberation is essentially *solipsistic*, limiting reasoning to, very roughly, discovering the most appealing means to one's ends and adjudicating conflicts between them, reducing reasons merely to considerations that help guide us to achieve our goals.

Conversely, the Rawlsian/Korsgaardian conception of reasoning rests on the claim that reasons are *systematically justifiable*, that is, one's claims to reasons are in-principle justifiable to all agents in rational equilibrium with their claims. As a result, reasons are not mere maps to the proper means for our ends, but tokens in the public exchange of interpersonal justification, in public reasoning. Because our episodes of reasoning can encounter the claims and complaints of other agents, it can lead us not merely to reflect on which ends we desire, but also their desirability, in apparent contrast with the Williamsian conception. Reflecting on the desirability of our ends allows us to "create new and destroy old underived desires by trying to come up with a systematically justifiable set of desires" (161), desires that, ultimately, converge.

Smith's arguments for siding against Hume and with Kant on the relationship between rationality and reasons continue to provoke important reflection on some of philosophy's most enduring themes. This is partly because convergence may seem like the *consequence* of adopting the idea that *systematic justification* is central to rationality, but it's *actually* (or, at least, *also*) the consequence of one of Smith's platitudes about desirability. After all, what's desirable that we do is an *agent-invariant* fact. Consequently, (1) implies that what we would desire to do if we were fully rational is an agent-invariant fact, meaning that what's advised by our ideal selves is agent-invariant.

This platitude about the relationship between desirability and reasons for action not only helps anchor the dispute between Williams and Smith concerning the possible scope of reasoned change in our desires, it also clarifies Smith's claims when he writes,

> Does our concept of a normative reason presuppose that there will, or alternatively that there will not, be a convergence in the desires that we would have under conditions of full rationality? If it presupposes that there will not be such a convergence, then our concept of a normative reason is relative. If it presupposes instead that there will be such a convergence, then our concept of a normative reason is, by contrast, non-relative. (p.166)

In short, because ideal advice doesn't vary between agents — initial differences are ironed out by systematically justify our desires — and because our normative reasons correspond to what's ideally advised, (1) and (2) imply:

3. What we have normative reason to do is what it is desirable that we do.

This claim is, I'll argue, a source of deep tension in Smith's account.

Relativism

Despite his commitment to a non-relative conception of reasons, Smith acknowledges that reasons can nevertheless be agent-relative in a couple of different senses. Indeed, *The Moral Problem* distinguishes three senses of relativity involved in claims about reasons, two of which Smith embraces and one of which he rejects. One form, which he attributes to Williams, relativizes normative concepts themselves, concepts such as DESIRABLE or RATIONALLY JUSTIFIABLE. We might analogize this relativity to etiquette: different groups have different codes of etiquette. When a member of one groups speaks in an unqualified, autocentric sense of what's polite, they're expressing claims using their culture or group's particular sense of 'polite', which we can mark with a subscript *a la* Smith (1994: 167), distinguishing, for example, POLITE_{A} from POLITE_{B} and so on. Without an independent desire not to offend, the standards of politeness of one group have no special bearing on the conduct of members of another group. For example, confronting some member of group A with the charge that their conduct is rude_{B} is irrelevant to them, at least absent an independent desire not to offend. The remark has merely sociological import. Asking the member of A to change their conduct on the grounds that it is rude_{B} would be merely to browbeat them into behaving differently.

In this sense, we might say that a *concept* of what's polite or rude is relative to a group; its "prescriptivity", "normative force", "bindingness", "rational enjoinment of motivation", (and so on) depends on accepting the

relevant group's code of etiquette. As a consequence, when people engage, unknowingly, in cross-cultural disagreements about whether someone's conduct is rude, they can end up talking past each other. For example, just as people can end up in an apparent disagreement about whether someone is tall when one means tall (for a philosopher) and another means tall (for a professional basketball player), the same can happen when one person asserts that the conduct was rude_A and the other denies that it is rude_B.

On Smith's characterization, Williams' conception of what's desirable or rationally justifiable functions similarly to how POLITE functions, except that those concepts are tethered to the pursuit of ends we have or could have rather than a group's standards of etiquette. Differences in each person's ends underlie differences in each's concept of DESIRABLE or RATIONALLY JUSTIFIABLE. Agents avoid talking past each other when discussing what's desirable or justifiable only to the degree that their ends overlap.

Smith contrasts this form of relativity with two others. Both correspond to explicit elements of canonical reason attributions such as that Nicole is hungry is a reason for her to go to the pizza parlour if she's nearby. We can represent this idea more generally by stipulating that a fully explicit reason attribution holds that the consideration that P is a reason for agent A to pursue option φ in circumstances C, summarizing this idea with the reason relation R<P, A, φ, C>.[3]

Smith recognizes variability between agents about whether a certain consideration supports pursuing a certain option in certain circumstances. Sometimes a consideration is a reason for some but not others to φ in C, *i.e.*, *relative* to some but not others. This contrasts with the previous kind of relativity because, regardless of what's true of us antecedently, when you and I have different reasons to do things, we have reasons in precisely the same sense. This isn't true of politeness, for example; we might embrace different senses of 'polite'.

Similarly, Smith recognizes variability in the circumstances under which there's a reason for someone to do something. The fact that there's beer at the store is a reason for you to go when you want beer but not when you don't. We can accommodate this by including preferences, such as a preference for beer, in the context mentioned as part of a fully explicit reason attribution. So just as a consideration can be a reason for one but not another in a given context, a consideration can be a reason for someone in one context but not in another.

Smith distinguishes these three forms of relativity to reject the first but embrace the latter two. It's natural to distinguish the first from the latter two since the latter two forms of relativity involve variation *within* the parameters of Smith's reason relation — *i.e.*, variation between *agents* and *contexts* —

3 *C.f.*, Scanlon (2014)

whereas variation in the first would involve variation between reason relations themselves or at least our concepts of those reasons, *i.e.*, variation between, for example, R_A<P, A, φ, C> and R_B<P, A, φ, C> and so on.

The critical difference between the kind of relativity that Smith accepts and the kind from Williams that he rejects is that only the former relativizes elements *within* a single common conception of a reason, shared by all practical agents. If agents share a common conception of reasons, they can agree or disagree about what there is reason for someone to do — they can engage in joint reasoning about a common subject matter. Conversely, Williams imputes different conceptions of a reason to different agents. It's hard to see how agents can discuss a common subject matter unless they share a common conception for disagreement about reasons between agents with different conceptions of reasons is as fruitless as disagreement about what's 'polite' between agents who accept different codes of etiquette.

Convergence in ideal desire is the product of systematically justifying our desires through the exchange of reasons. Williams-style relativism prevents convergence by preventing the exchange of opinions on reasons when agents' desires differ sufficiently. Consequently, because Smith embraces convergence, he must reject Williams-style relativism. But in striking contrast with his deep examination of the conflict between Williams-style relativism and convergence, Smith largely assumes that reasons' agent– and circumstance-relativity is compatible with convergence. For example, he writes,

> Suppose someone tells me that she has a reason to take a holiday and that I think I would have no reason to take a holiday in the circumstances she faces. Provided we have taken proper account of the *de se* considerations that might be relevant to her choice, and provided we have taken proper account of the way in which her preferences may constitute a relevant feature of her circumstances, it seems that I straightforwardly disagree with her about the rational justifiability of her taking a holiday in the circumstances she faces, a disagreement I can express by saying 'she thinks that there is a reason to take a holiday in her circumstances, but there is no such reason.' (p.171)

Whether agent-relativity, like Williams-style relativity, blocks convergence depends on whether, as Smith puts it, whether it is possible for the agent's perspective — so-called '*de se*' considerations — to be part of the public record, so that the exchange of reasons can take 'proper account' of them. But these considerations pose a larger challenge than Smith seems to recognize.

3. Convergence of What?

Smith argues that our desires converge under conditions of ideal rationality. Moreover, they converge on what is desirable for us to do. Thus, desirability is non-relative — it is "desirability *simpliciter*" (p.167) — so what is desirable

to one is desirable to all. But this position is subtly ambiguous. On the one hand, it's natural to suppose that ideal agents converge on *exactly the same desires*, in the sense of having the same states of mind. Yet, on the other hand, it's equally attractive to suppose that ideal agents converge on desiring the exact same objects of desire, understood as the intentional objects of those attitudes. *The Moral Problem* does not resolve this ambiguity. But regardless of how it's resolved, convergence in ideal desire is hard to reconcile with the book's broader aims.

For example, if we're running a race, and you want the trophy and I want the trophy, we want the same thing — the trophy. By the same token, it may seem that when you want to win the race and I want to win the race, we want the same thing: winning the race. But the two pairs of desires differ grammatically. The first pair uses a noun phrase to denote the content of the want, *viz.*, 'the trophy'. The second pair uses a non-finite clause to denote the content of want, *viz.*, 'to win'. 'Want' in this second pair is typically taken to denote a *propositional attitude*, and if so, 'to win' must denote a proposition.[4] So 'Jim wants to win the race' is typically read as expressing the thought that Jim wants that Jim himself wins the race, where the slightly infelicitous expression 'Jim himself' marks that the proposition denoted by 'that Jim himself wins the race' is a first-personal proposition of the kind normally expressed using 'I', such as the thought that I want to win (perhaps as thought by Jim).

When I think that I'm a philosopher and you think that you're a philosopher (in the first-personal way), there's a shallow sense in which we have the same thought — roughly akin to our sameness of thought when I think that *he's* the tallest person in the room (mentally ostending to the man on our left) and you think that *he's* the tallest person in the room (mentally ostending to a different man on our right). But it should be clear that these pairs of thoughts differ in an especially critical way: they are *about* different people, so they have different truth conditions. Said differently, while the pairs of thoughts may have the same (following David Kaplan (1989)) kind of *character*, the thoughts have different *contents*. As such, they count as different thoughts, despite a shallow resemblance in character.

Likewise, when you want to win the race and I want to win the race, our wants have a similar character. But they are different wants for they involve different contents (one involving *me* winning and another involving *you* winning). Because of this, the two wants are satisfied by different events: one of *my* winning and another of *your* winning. So while there's a shallow sense in which we want the same thing, *viz.*, to win, there's a much deeper and more central sense in which we want different things. I want the event where I win and you want the distinct, incompatible event where you win.[5]

---

4 See Chierchia (1990), for example, and Moltmann (2005) for elaboration.

5 I am arguing that individuals share fewer desires than it might seem, at first glance. In particular, whenever our desires involve first-personal contents, agents' desires diverge. However, this divergence is even more dramatic than may be initially supposed. It

Quite plausibly, adopting conditions of ideal rationality wouldn't erase my sense of self or my knowledge of my identity — indeed, this knowledge seems guaranteed by the idea that our ideal selves are ideal partly because they have "all relevant true beliefs" (p.156). As a consequence, it seems equally plausible that our ideal advisors will have some desires that involve themselves first-personally (or their less-than-fully rational advisees) — otherwise we could talk of single ideal advisor for all, occupying an entirely impersonal "standpoint of the universe", rather than convergence of individual advisors.

So it seems entirely plausible that convergence of desires in ideal advisors doesn't correspond to *identity* of desires in ideal agents. Convergence seemingly cannot consist of sharing the same desires; the desires of ideal advisors don't converge on the same *contents*.

Perhaps we might defend the claim that ideal desires converge in content by denying that desiring the desirable requires desires with *de se* contents. But this position is difficult to reconcile with competitive scenarios like the one above. For example, there are situations where it's desirable, at least seemingly, to win. When we both desire to win in these cases, we desire what's desirable. But these desires are covert *de se* desires. It would be convenient for this answer if the appearance of *de se* content in these desires were inessential, then we could dispense them with them without threatening convergence. But it isn't obvious how to offer an impersonal paraphrase of them, one lacking *de se* content. For example, when you and I both want to win, the surface grammar suggests that we want the same thing after all: the quality of winning or being the winner. But we don't merely want the exemplification of this property. It matters *who* does the exemplifying. If you become the winner, I won't get what I want. Rather, what I want is for *me* to exemplify the quality of winning. So long as something desirable is picked out by a 'to'-clause embedded under a pro-attitude, we cannot dispense with *de se* attitudes. Consequently, it seems that convergence cannot be convergence in the contents of our pro-attitudes.

Perhaps a more plausible candidate for convergence is desiring the same *things*, regardless of the varied ways in which those things figure in our minds to yield different contents. So if it's desirable that Simone Biles wins the gold medal, our ideal advisors will desire that Simone wins while ideal Simone will desire that *she herself* wins. Of course, these are different desires — the latter is *de se* but the former is not — yet they converge on the same event, *viz.*, the desirable event of Simone winning gold. According to this view, differences in how someone is presented in thought, at least in the sense that "everyone is presented to himself in a special and primitive way, in which he is presented

may be that *all* desires diverge in this sense. For example, Milona & Schroeder (2019) argue that desires reported using determiner-phrases such as 'the trophy' as in 'I want the trophy' are to be understood in terms of non-finite phrases — roughly, 'I want the trophy' expresses that I want to hold/own/win/etc. the trophy. If that's so, and if non-finite phrases embed the first-person perspective on those contexts, then overlap in our desires will plausibly be limited to desires with *explicit* overlap in content, such as when you and I want *us* to win. Examining these details takes us too far afield.

to no one else" (Frege 1956: 298), don't affect whether the desire is ideal or not. What matters instead is the object of the desire — in particular whether that object is desirable — not its content.

The claim that ideal desires converge on certain objects makes convergence impartial or agent-neutral, since ideal desires converge on something just when it is desirable simpliciter. This makes the claim appealing. That's because morality is importantly impartial in the broad sense that we're all moral equals. That thought suggests that whether we're one individual rather than another is not *in itself* a fact of moral import; it matters only insofar as it's connected to other facts of independent concern, such as whether we're virtuous or vicious, rational or not, etc. As Bernard Williams puts this idea in *Utilitarianism: For and Against*, if convergence is impersonal,

> Such a principle will claim that there can be no relevant difference from a moral point of view which consists just in the fact, not further explicable in general terms, that benefits or harms accrue to one person rather than to another – 'it's me' can never in itself be a morally comprehensible reason. (Williams 1973: 96).

But the idea that convergence is impersonal seems to conflict with Smith's apparent commitment to agent-relative reasons and, in particular, to a class of reasons that he calls "*de se* reasons".

De Se Reasons

First-personal thoughts figure prominently in *The Moral Problem*'s conception of normative reasons. Smith implicitly relies on them — calling them '*de se*' thoughts, following Lewis — to support his idea that agents and circumstances are distinct parameters or dimensions of variation between reasons. But it's not immediately clear why he includes both parameters in his conception of reasons. Since every agent is located in a circumstance, it might be tempting to subsume the agent parameter to the circumstance parameter. For example, when there's dancing at the party, there's a reason for Ronnie, who loves dancing, to go but not for Bradley, who hates dancing, to go. The prepositional phrases 'for Ronnie' and 'for Bradley' suggests that we are dealing with agent-relative reasons. But Ronnie and Bradley's preferences explain the differences in their reasons, and one's preferences are part of one's circumstances. So we can capture variation in Ronnie and Bradley's reasons with the idea that there's reason for *anyone* to go to the party in the circumstance that they like dancing.

This manoeuvre, which Schroeder (2007a: 43-56) calls 'the Standard Model', is especially attractive if we hold that moral reasons are agent-neutral reasons and that promises yield moral reasons to do what's promised. Just as you and I can differ in whether we like to dance, and so too differ in our reasons for going to the party, you and I can differ in whether we've promised to go to the party. If you've promised and I haven't, you have a reason to go and I haven't. This makes it seem as though you have only an *agent-relative*

reason to go, which contradicts the idea that promising gives an agent-neutral reason to go because promises give moral reasons.

But whether or not one has promised to go to the party is part of one's circumstances, or so we can assume. So we can explain the difference in our reasons to go to the party with the difference in our circumstances, not with the idea that promises yield agent-relative reasons. Rather they yield *agent-neutral* reasons of the from that *anyone* has a reason to do what they've promised. Indeed, it might seem tempting to suppose that *all* variation between reasons is variation in circumstances, that agent-relative reasons are grammatical illusion. Indeed, Howard and Schroeder (2024: 51-2) characterize Smith's position in those terms. While this characterization captures part of the spirit of Smith's approach, it seems to contradict an important discussion of agent-relativity in *The Moral Problem*:

> Sometimes what we have in mind when we say "That may be a reason for you, but it isn't for me" is that the considerations that rationally justify our choices are, to use Parfit's terms, *agent-relative*, rather than *agent-neutral* (Parfit, 1984). Suppose you are standing on a beach. Two people are drowning to your left and one is drowning to your right. You can either swim left and save two, in which case the one on the right will drown, or you can swim right and save one, in which case the two on the left will drown. You decide to swim right and save the one and you justify your choice by saying "The one on the right is my child, whereas the two on the left are perfect strangers to me". [...] [T]here are both *de dicto* and *de se* normative reasons. We can each express the content of the *de dicto* reason relevant in this case by using the words "There is a reason to save people quite generally" and we can each express the content of the *de se* reason by using the words "There is a reason to save my child in particular". (pp. 168-9)

It's natural to see the fact that it's *my* child is a reason only for some, and not for all, to prefer that I save my child rather than the two perfect strangers — for example, presumably the strangers' parents feel very differently about the matter. So it's an agent-relative reason. Smith's observation that this reason is *de se* (it's *my* child, after all) offers a compelling explanation of why the reason is genuinely agent-relative, not just covertly agent-neutral through variation in the circumstance parameter. The explanation comprises three claims:

A. There's a motivational constraint on reasons such that P is a reason for S to φ only if S can φ for P.
B. First-personal thoughts are *private* in the sense that a first-personal thought is thinkable only by its subject.
C. Acting for a reason requires thinking the thought corresponding to the reason.

Because *de se* reasons correspond to first-personal thoughts, only the subjects of *de se* reasons can act for them, from B and C. If that's right, then *de se* reasons are reasons only for the agent that can think the corresponding thoughts (from

A). So *de se* reasons are genuinely agent-relative reasons, not just covertly agent-neutral reasons, seemingly *contra* Howard and Schroeder (2024).

Yet there's a deep tension between accepting both the thesis of impersonal convergence and genuinely agent-relative *de se* reasons. If ideal desires converge by converging on the same set of objects (understood broadly to include events, states of the world, possibilities, and so on), then there's little room for *de se* reasons to affect what's desirable where a connection to what's desireable is characteristic of reasons given the claim established above that:

4. What we have normative reason to do is what it is desirable that we do.

To illustrate this, return to Smith's case and modify it slightly so that you apprehend that both your child and a stranger's two children are in urgent mortal need of help. You must choose whom to save; you cannot save all. Smith allows that 'It's *my* child!' offers a genuine agent-relative reason for you to save them rather than the two strangers. Of course there's a countervailing agent-neutral *de dicto* reason to save the strangers flowing from the fact that "there's reason to save people quite generally" (p.169). Now it's a possibility that *de re* reasons are inconsequential to the systematic justification of our desires. They may not figure *at all* in systematic justification or they may be consistently defeated by countervailing considerations.

But it will seem equally desirable to many that you save your child rather than the two strangers and, indeed, strange if you don't not feel the pull of that systematically justifiable desire. If we accept this, then we allow that it is desirable that you save your child in virtue of your *de se* reason. So we might be attracted to the general claim that:

5. It is desirable that parents save their children rather than strangers' children.

Of course, we can find problematic exceptions to this general claim ("What if your child is baby Hitler!?"). But (5) should be read as a generic such as "Lions have manes" and generics characteristically admit counter instances. As such, if we are tempted by the idea that there are *de se* reasons such as those described by Smith, and if we accept the connection between reasons and desirability in (4), then we should be equally tempted by (5), which captures some aspect of a parent's special obligation to their child.

This aspect is, however, largely peripheral to how we think of parental duties, at least to how we think of them as prompted by Smith's case. Recall that what's desirable is an agent-invariant matter; desirability is "desirability *simpliciter*" (p.167). So what (5) expresses is the public benefit of parents caring for their children. It's the kind of impersonal benefit pursued by state policies of paid parental leave whose justification is the social benefit that those policies provide to all, *viz*., the idea that we are all better off in a society that supports and facilitates, to at least some degree, the fulfilment of parental

duties. But we don't need to invoke *de se* reasons to explain social benefits. Reasons voiced by publicly accessible facts are entirely adequate.

As a consequence, (5) expresses neither the special duties enjoined by a parent's special relationship to their child nor how that relationship generates a special reason for that parent and for no one else. Cases where parents have conflicting — that is, mutually exclusive — interests demonstrate this shortcoming. Suppose that we add to the scenario above that the other child's parent arrives at the scene to see their child drowning and suppose further that only one child can be saved, say, because there's only one life-preserver, which is necessary for rescuing either child. Each parent in this situation has a *de se* reason to save their child rooted in the thought 'it's *my* child' as thought by each. These reasons seem to support conflicting outcomes, with your reason supporting the outcome where your child is saved rather than the other child and *mutatis mutandis* for the other parent. Thus, given (4), each of these *de se* reasons implies, respectively:

6. It is desirable that your child is saved rather than the other child.
7. It is desirable that the other child is saved rather than your child.

But the only way for both claims to be true is if it's equally desirable *simpliciter* that either child is saved. Now this position perhaps reflects the judgment of a bystander, for whom the loss of either child is equally tragic. But it does not reflect the judgment of either parent, for whom the loss of *their* child is a special tragedy. It's presumably the latter's perspective that *de se* reasons aim to capture. We could discount the parent's perspective on the grounds that they're particularly biased, and so deceived in some broad sense, about the desirability of their child being saved. This simplifies axiological matters considerably. But then *de se* reasons serve little purpose save perhaps as distinguished motivating reasons, underwriting the parental virtue but not the rightness of each parent's attempted rescue. If that's right, *de se reasons* do not merit special mention in the context of Chapter 5.9's discussion of *normative* reasons; *de se* reasons would seem to be normatively epiphenomenal concerning desirability. And, further, unless *de se* reasons are genuine normative reasons, then the distinction between agent-relative and agent-neutral reasons collapses, just as Howard and Schroeder suggest.

Likewise, some might insist that the process of systematic justification would purge parents of their biased preferences for their children; but this is just to deny that *de se* reasons are real reasons, since *de se* reasons are the ones that survive the process of systematic justification. So it cannot be that it's irrational to prefer the flourishing of one's child, since that would imply that there's no *de se* reason to prefer that flourishing where Smith explicitly recognizes one.

Smith is thus trapped between a rock and a hard place. He has good cause for accepting that *de se* reasons are genuine, agent-relative normative reasons, both from the idea that parents have special reasons to be partial

to their children and from the desire to distinguish genuine agent-relativity in reasons from relativity to an agent's circumstances. However the agent-relativity of *de se* reasons conflicts with the convergence of ideal desire for those desires converge on what's impersonally, agent-neutrally desirable. Consequently, Smith's position seems to face a version of the objection that G. E. Moore (1903: §59) wields against egoism when he argues that combining it with consequentialism yields an "absolute contradiction".

4. Good Aims

One response to this challenge is to replace the idea that the desires of ideal advisors converge on what's desirable with the idea that they converge on what's fittingly desired, and allow for variation in what each person('s ideal advisor) fittingly desires. Smith offers an extensive discussion of how fitting desires can vary between agents in Smith (2003) and Smith (2009). I'll offer a brief criticism of this approach but since my aim here is to offer an alternative I won't claim that this criticism is decisive or unanswerable, only that it gives some reason for looking to the alternative that I advocate.

Smith's approach, especially as voiced in Smith (2009), distinguishes between what he calls *evaluator-relative* and *non-evaluator-relative value-making features*. For our purposes, the indexical or first-personal features that underlie *de se* reasons are paradigmatic evaluator-relative value-making features. Publicly accessible facts, such as that a child is drowning or that people are in need, are paradigmatic non-evaluator-relative value-making features. Given that *de se* reasons are private and inaccessible to others, variation in our *de se* reasons can underlie variation in what's fitting for each of us to prefer (*c.f.*, Ewing (1947; 1959)) or what we have reason to prefer (*c.f.*, Scanlon (1998)) or what merits our approval (*c.f.*, McDowell (1985); Wiggins (1987)). Evaluator-relative value-making features allow us to construct an *agent-relative* sense of desirable — a conception of agent-relative value — that, for example, eliminates the conflict between (6) and (7) while ratifying the partial preferences of each parent.

Smith defends these claims against a well-known criticism from Mark Schroeder (2007b) that agent-relative value is an unsuitable basis for consequentialism. Schroeder's criticism can be put simply:[6]

> P1. Agent-relative consequentialism tells each agent to pursue what has the highest agent-relative value, relative to them.
>
> P2. Our conception of agent-relative value is artificial, a "theoretical posit".
>
> P3. If P2 is true, then it's implausible that, for each agent, there is a reason for them to pursue what has the highest agent-relative value, relative to them.

6 I discuss Schroeder's argument and expand the criticism below in Howard (ms).

C. Therefore, it's implausible that there's reason to obey agent-relative consequentialism.

Smith (2009) explicitly argues (pp. 268-71) that his appeal to evaluator-relative value-making features rebuts P2 in Schroeder's argument.[7] It seems that being *my* child, relative to me, is a distinct value-making feature from being Nathan Howard's child.[8] Only I can prefer or disprefer outcomes based on whether they involve *my* child; moreover, in many situations it can be fitting for me to prefer my child's flourishing to another child's even though, for all intents and purposes, each child's flourishing is equally good, in an impartial sense. Consequently, the term 'agent-relative value' tracks the property that an outcome has when it's fitting for one agent but not another to desire or prefer it. (We can put the same point in terms of 'merit' or 'possessed reasons', if we prefer.) Nothing about this notion is artificial for its conceptual ingredients are entirely commonsense.

This response wins the battle against P2 but it loses the broader war concerning Schroeder's conclusion, C. For suppose that there are indeed *de se* or other agent-relative reasons to prefer that, for example, my child flourishes rather than some other child. If these reasons are to play their role in *The Moral Problem* or in consequentialist theory writ large, they must be properly connected to reasons for action. *The Moral Problem* entails:

8. What we have normative reason to do is what it is desirable that we do.

The corresponding claim concerning agent-relative reasons is:

9. What each has normative reason to do is what it is desirable-relative-to-them that they do.

But (9) is dubious in a way that (8) is not. Here is an argument against (9):

10. All agent-relative reasons for desire are wrong-kind reasons for desire.
11. Wrong-kind reasons for desire don't generally transmit to reasons for action.
12. (9) implies that agent-relative reasons for desire generally transmit to reasons for action.
13. Therefore, (9) is false.

(10) follows from the definition of a wrong-kind reason for preference. A reason for preference is of the wrong kind if it's excluded from the buck-passing analysis of betterness:

7 See also Hammerton (2020).

8 I think treating indexical features as features of outcomes *already* commits Smith to the proposal that I go on to suggest, but this depends on additional commitments that Smith does not explicitly adopt.

> **Buck-Passing Analysis of Betterness**: One thing is better than another else just when there is sufficient reason for anyone to prefer the first to the second.

De se reasons are not reasons for anyone; they're reasons only for the subject of the consideration that gives the reason. Consequently, they must be excluded from the analysis at the risk of falsifying it. After all, the strength of my *de se* reasons to prefer my child's flourishing mean that I have sufficient reason to prefer that my child flourishes rather than another child. But my child's flourishing is no better than that child's flourishing, or so we can assume.

(11) is simply a more-or-less banal observation about wrong-kind reasons for preference and desire in general. Threats and bribes to prefer things provide paradigmatic instances of wrong-kind reasons — what Jonathan Way calls 'incentives'. Suppose an evil demon threatens you with harm unless you prefer to sip from one saucer of mud rather than another equivalent saucer of mud. Plausibly, though controversially, the incentive gives a reason to prefer the first saucer to the second. But it doesn't also give you a reason to sip from the first saucer. After all, you avoid harm by adjusting your preferences, not by sipping. (12) makes (9) explicit, so rejecting it is out.

If the argument is sound, it shows that although Smith has articulated a non-artificial ("organic"?) conception of relative value, it's not a conception that's properly connected to action, so it's not a suitable basis for consequentialism, just as Schroeder's conclusion asserts. I stress that much more needs to be said to establish the argument's soundness and to assess the moves available to Smith and countermoves available to critics, although I do this elsewhere.[9] I mention the argument here only because it motivates the alternative method of reconciling *de se* reasons and desirability that I prefer.

In Howard (2022), I argue that consequentialists who recommend performing the action with the highest *expected* value, so-called "subjective consequentialists", employ a conception of outcomes that allows us to reconcile agent-centered prohibitions on certain actions — such as killing or lying, perhaps — with the idea that we should pursue what's agent-neutrally best. That conception of outcomes accommodates agent-centred prohibitions because it individuates outcomes more finely than states of the world. For example, necessarily, every situation where a doctor prescribes paracetamol is a situation where they prescribe acetaminophen — they're different names for the same drug. But in situations where the doctor is misinformed and falsely associates paracetamol with risks that she doesn't associate with acetaminophen, prescribing paracetamol can have higher expected value for her than prescribing acetaminophen, despite the fact that every event of prescribing one drug is identical to an event of prescribing the other. In short, when an agent is ignorant that two concepts are co-intensional — when she is

9 Howard (ms).

"Frege-puzzled" — she can divide value differently over two outcomes when there is only one state of the world.

First-personal aims, such as the aim where *I*, thought of first-personally, receive a benefit are just a special case of this phenomenon. In a situation where an egoist is Frege-puzzled about their identity, they can prefer that they *themselves* receive a benefit rather than some stranger even if, unbeknownst to the egoist, they're the stranger. And if we're egoists ourselves, we may even find the egoist's preference fitting. So fitting preferences can depend on Fregean differences.

It should be plain from the discussion above how we can accommodate these judgments if we take desirability to be an agent-relative concept. But as Smith stresses in *The Moral Problem*, desirability is not an agent-relative concept and Schroeder gives us further cause to doubt the suitability of agent-relative value as a foundation for ethics. So we must look elsewhere to accommodate them.

Fortunately, if we accept both that (a) *de se* aims, such as that *my* child is rescued, are distinct from impersonal aims, such as that Nathan Howard's child is rescued, and that (b) their constituent *de se* thoughts are private, we have all we need to resolve the tension in *The Moral Problem* identified above and reconcile *de se* reasons with desirability. The key insight is that — and I fully recognize that this claim initially seems odd though I will go on to argue that it is not — certain *thoughts* or *aims* are especially worthy of desire. For example, suppose that you come across your child, who is drowning. However, owing to a trick of the light on the waves, it seems as though there are *two* children drowning, yours and a stranger's. Although the aim of saving *your* child and the aim of saving "the stranger's child" correspond to exactly the same event, *viz*., saving your child, certainly the first aim is more worthy of your desire than the second. And, I suggest, it is worthy *simpliciter* of that desire; we don't need to add a *second* relativization, such that the aim of saving *your* child is worthy of desire *relative to you*. The aim itself, of saving *your* child, is already agent-relative, owing to its essential inclusion of *de se* content. It is what I call a *de se* aim.

So this is my proposal for reconciling *de se* reasons and desirability. *De se* reasons support the adoption of *de se* aims, which, when they involve the flourishing of our nearest and dearest, are particularly valuable, indeed, more valuable than co-referring aims, which involve those very individuals' flourishing but represented impersonally. When we bear the kind of intimate relations with someone that give rise to special *de se* reasons, we are also in a position to adopt special *de se* aims, which we have particularly strong cause to pursue, as those aims are favoured by *de se* and *de dicto* reasons alike.

This proposal is not vulnerable to the kind of worries that attend agent-relative normative concepts, such as the ones mentioned above, voiced both by Smith, discussing Williams, and by Schroeder. After all, *de se* aims are

agent-relative in just the same way that *de se* reasons are. Just as *de se* reasons *support* certain aims in a non-agent-relative sense ("our concept of a reason is stubbornly non-relative" (p. 172)), *de se* aims are worthy of that support in an equally non-relative sense. In both cases, it's the objects themselves that are relativized to agents, not the normative qualities that they exemplify.

Colleagues typically baulk at the same point when I describe this approach, which sometimes allows evaluative expressions — "it is desirable (to such-and-such degree) that", "is better than", etc. — to behave like hyperintensional operators, akin to attitude verbs such as 'believes' or 'desires'. Just as Jim can think that Superman flies without thinking that Clark Kent flies, so too can it be *good* that my daughter flourishes without it being equally good that that child (who happens to be my daughter) flourishes. And this, my colleagues insist, is weird — it sounds like an attitudinal content, such as a proposition, is good. However, none of this amounts to an objection.

First, goodness and desire have long been linked. But some misunderstand this link as it relates to relative value. They see ideal desires as an attractive perspective on relative value because desires are relations between agents and the objects that they desire, allowing my ideal desires to differ from your ideal desires in the manner that agent-relativity requires. But I think that connection is less explanatory than the kind of agent-relativity exhibited by *de se* reasons and *de se* aims. According to these concepts, what matters for accommodating agent-relativity is not the fact that desires involve relations to agents but the fact that desires are sensitive to the different ways that aims can be presented and particularly to the difference between an aim being presented personally or impersonally.

This link between value and desire suggests several compelling arguments for interpreting "it is good that" in the hyperintensional manner described above when applied to aims. Here's a particularly straightforward one. Many, but most prominently Mill, think that being good just is being desirable. Further, we can analyze what's desirable as what's fittingly desired, echoing Ewing and his followers. If all desires are hyperintensional operators (assuming that desire is a propositional attitude), so too are fitting desires as a special case. If goodness just is what's fittingly desired, then "it is good that" is therefore also a hyperintensional operator. But this shouldn't surprise us at all: we're supposing that "it is good that" denotes the same quality as "it is fitting to desire that".

Second, I take it that what's truly controversial is not the semantic claim that "is good" is a predicate of contents, resembling "desires", but the metaphysical claim that goodness is a property of the contents of desires. I've only defended the first claim; it does not entail the second except given dubious assumptions about semantic descent. For example, the same inference via semantic descent fails in the case of desire: we can't conclude from the fact that "Jane desires" in "Jane desires that there is world peace" is

a predicate of propositions that the proposition that there is world peace is what Jane desires. Likewise, we can't conclude from the fact that "is good" is a predicate of contents that contents are what's good.

To elaborate, suppose that I want to meet Superman more than I want to meet Clark Kent. In that case, that I meet Superman is more desired-by-me than that I meet Clark Kent. But that isn't *made true* by the proposition that I meet Superman exemplifying the property of being desired-by-me to a greater degree than the other proposition. Rather, it is made true by something else — in this case, my psychological state. Likewise, when some indexical prospect is better or worse than its non-indexical counterpart that need not be true in virtue of the fact that the degree of value exemplified by the indexical prospect differs from the value of the non-indexical one. It could be made true by something else, as in the case of "desires" — perhaps facts about which desires are fitting, as above. Consequently, the conclusion that aims are value-bearers is optional.

Nevertheless, despite these remarks, the idea that aims are value-bearers is less offensive to common sense than it might seem. Axiologists theorize about the property that an object has when it's worthy of non-instrumental pursuit, promotion, admiration, respect, honour, etc. Certain aims are especially worthy of non-instrumental pursuit, such as the aim of benefitting your child. Moreover, aims, in at least one sense, are individuated in the manner of propositions. You can aim to ease someone's pain without aiming to suppress the firing of their c-fibers even if, necessarily, easing one is suppressing the other. So we can say that some aims are more worthy of pursuit than others, and that's just to say that some aims are better than others.

A helpful referee raises the following concern: if *de se* aims diverge between agents, then are they compatible with the universality of desirability? If not, then it would seem that good *de se* aims are only agent-relatively, and not agent-neutrally, good, making them vulnerable to precisely the worries from Schroeder and others that I hope to avoid.

Let me be explicit: the aim of saving *my* child, as thought by me, is an abstract object — enjoying whichever ontological status we attribute to propositions more generally — that is agent-neutrally worthy of desire, that is, good *simpliciter*. In that sense, the aim is universally worthy of pursuit or desire.

However, because the aim is *de se*, it is *private* in the manner described by Frege. As a matter of metaphysical fact, only *I* can adopt the aim, so only *I* have a special responsibility to pursue it in the manner consistent with my special obligation to my daughter. Thus, the privacy of *de se* aims reconciles the *agent-neutral* character of *de se* aims' value with the *agent-relative* character of the duties and permissions to which those aims give rise.

To illustrate, suppose tragedy strikes. Both your child and mine are drowning. Further, there's only one life-preserver with which to save them. If

we have special obligations to our children, then we should pursue conflicting actions: I have a reason to grab the life preserver to save my child rather than yours and *mutatis mutandis* for you, where we cannot "share" the preserver to save both in the relevant sense. These special obligations look incompatible with (3), Smith's idea that "What we have normative reason to do is what it is desirable that we do" for *either* it is equally desirable that either child is saved, in which case neither of us has a special reason to save our child, or it is more desirable to save one child rather than the other, in which case one of us lacks a reason to save our child.

But according to the proposal above, our differing special obligations are tied to subtly different *de se* aims: *my* aim of saving *my* child, and *your* aim of saving *yours*. The special (agent-neutral) value of these *de se* aims accounts for our special obligations to our children. Moreover, since you cannot adopt my *de se* aims, for they are private in the Fregean sense, the special value of saving *my* child does not interfere with your special reasons for saving your child. Consequently, we can trace the agent-relative import of *de se* reasons to the privacy of good *de se* aims.

I think significant resistance to the idea that some aims are better than others, in the sense of aim that corresponds to contents or propositions, can be attributed to a misunderstanding. When I claim that there are good aims, and by this I mean that there are good propositions, some are liable to understand me as saying that some aims are *attributively good* propositions — or perhaps, adopting an idea popularized by Peter Geach, that there are aims that fulfil the function of propositions especially well.[10] But of course, I mean 'good' in the sense of predicatively good. When I say that there are good aims, I mean only that some aims are particularly worthy of pursuit, where pursuing a given aim doesn't entail pursuing every aim that's co-intensive with the first.

5. Conclusion

Although its topic is complex, the structure of the preceding discussion is fairly simple. I began by revisiting Smith's threefold distinction in kinds of moral relativity: concept relativity, circumstantial relativity, and what, following Parfit (1984), is properly called agent-relativity. I then reconstructed an argument for distinguishing the latter two forms of relativity based on Smith's discussion of *de se* reasons. I then argued that accepting *de se* reasons creates a tension with a central thesis of *The Moral Problem*, the thesis that ideal desires converge on the desirable. The tension is easy to appreciate: since the desirable is agent-neutral and *de se* reasons are agent-relative, they tend to pull in opposite directions. For example, it seems that the latter but not the former allows for personal relationships, such as parent to child, to have a private evaluative significance that they lack publicly. I then argued

10 Geach (1956).

that the solution to this tension is not to replace claims about desirability with the notion of agent-relative value, which Smith explores in more recent work. Rather, the solution is to extend the distinction between *de dicto* and *de re* reasons to some value bearers themselves, what I called aims, which I argued is less alien than it might seem at first.

References

Bukoski, M. (2016). A Critique of Smith's Constitutivism. *Ethics*, 127(1): 116–146.

Chierchia, G. (1990). Anaphora and Attitudes. In *Language in Action*, R. Bartsch et al. Dordrecht: Foris Publications.

Ewing, A. C. (1947). *The Definition of Good*. London: MacMillan.

Ewing, A. C. (1959). *Second Thoughts in Moral Philosophy*. London: Routledge and Kegan Paul.

Geach, P. T. (1956). Good and Evil. *Analysis*, 17(2): 33–42.

Hammerton, M. (2020). Relativized Rankings. *The Oxford Handbook of Consequentialism*. Oxford University Press.

Howard, N. R. (manuscript). Relative Value and the Teleological Conception of Practical Reasons.

Howard, N. R. (2022). Consequentialism and the Agent's Point of View. *Ethics*, 132 (4): 787–816.

Howard, N. R. & Schroeder, M. (2024). *The Fundamentals of Reasons*. Oxford University Press.

Kaplan, D. (1979). On the Logic of Demonstratives. *Journal of Philosophical Logic*, 8: 81–98.

McDowell, J. (1985). Values and Secondary Qualities. In Ted Honderich (ed.) *Morality and Objectivity* (London: Routledge and Kegan Paul), pp. 110–29.

Milona, M. & Schroeder, M. (2019). Desiring Under the Proper Guise. *Oxford Studies in Metaethics*. Oxford University Press.

Moltmann, F. (2005). Generic One, Arbitrary PRO, and the First Person. *Natural Language Semantics*, 14: 257–281.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press.

Scanlon, T. M. (2014). *Being Realistic About Reasons*. Harvard University Press.

Schroeder, M. (2007a). *Slaves of the Passions*. Oxford University Press.

Schroeder, M. (2007b). Teleology, Agent-Relative Value, and 'Good'. *Ethics*, 117(2): 265–295.

Smart, J. J. C & Williams, B. (1974). *Utilitarianism: For and Against*. Cambridge University Press.

Smith, M. (1994). *The Moral Problem*. Blackwell Publishing.

Smith, M. (1997). In Defense of The Moral Problem: A Reply to Brink, Copp, and Sayre-McCord. *Ethics*, 108: 84–119.

Smith, M. (2003). Neutral and Relative Value After Moore. *Ethics*, 113(3): 576–598.

Smith, M. (2009). Two Kinds of Consequentialism. *Philosophical Issues*, 19: 257–272.

Wiggins, D. (1987). A Sensible Subjectivism. In *Needs, Values, Truth* (Oxford: Basil Blackwell).

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

1

FILOZOFSKI godišnjak = Belgrade philosophical annual / editor Voin Milevski. – God. 1, br. 1 (1988)– . – Belgrade : Institute of Philosophy, Faculty of Philosophy, 1988– (Belgrade : Službeni glasnik). - 24 cm

Polugodišnje. – Glavni stvarni naslov od br. 28 (2015) Belgrade philosophical annual. – Tekst na engl. jeziku.

ISSN 0353-3891 = Filozofski godišnjak

COBISS.SR-ID 15073792