Published in *The Future of Normativity*. Edited by Simon Kirchin.
Oxford: Oxford University Press. 2025. Pp. 341-366

THE ETHICS OF AGENCY: AN ALTERNATIVE APPROACH TO NORMATIVE ETHICS

Michael Smith

Is there a way to approach problems in normative ethics that is different from the ways in which they have been approached in the recent past? In what follows I argue that there is. To anticipate, the alternative supports an ethics of agency that affirms a pair of moral principles that need to be weighed against each other, one similar to a deontological principle and the other similar to a consequentialist principle. However, the similarity with familiar approaches ends there.

The derivation of the two moral principles begins with a novel solution to an old metaethical problem. As we will see, beginning in this way enables us to avoid common pitfalls when we go on to do normative ethics; it suggests new perspectives on some familiar problems in normative ethics; and it makes salient new problems in normative ethics for us to solve. Whereas existing forms of deontology and consequentialism are moral theories, the alternative suggests a more comprehensive view of normative ethics as including a large non-moral component and when the components conflict, it tells us how to balance not just moral considerations against each other, but also moral and non-moral considerations against each other.

Before spelling out the ethics of agency, we need to remind ourselves how and why problems in normative ethics have been approached in the ways in which they have recently. As we will see, understanding what happened in the first half of the twentieth century is especially important. In response to what happened then, one of the now-familiar approaches began mid-century with a bad solution to a metaethical problem, and the other began at much the same time by turning its back on metaethics altogether. It should therefore come as no surprise that an alternative approach might begin with a new solution to an old metaethical problem.

1. Ethics in the first half of the twentieth century

Our brief history begins with the rise of logical positivism in the mid 1920s and the subsequent publication of A. J. Ayers *Language, Truth, and Logic* (1936). These developments led many philosophers to think that the rise of meta-ethics at the beginning of the 1900s had led theorists astray. The culprit was G. E. Moore who had modelled his approach to ethics in *Principia Ethica* (1903)—non-naturalistic realism in meta-ethics and consequentialism in normative ethics—on Newton's approach to physics and mathematics.

The main problem with Moore's approach, according to the positivists, is that it falsely assumed realism—that there exists a domain of moral facts. Given this assumption, Moore thought the first task of the ethicist is to solve the problem in metaethics of giving an account of the nature of those facts. The exclusive options were supposed to be that moral facts are either amenable to scientific study (that is, they are *naturalistic* facts) or they are not amenable to such study (that is, they are *non-naturalistic* facts). Moore's major contribution was the infamous Open Question Argument that he thought demolished the former option, leaving us with the latter.

The problem with this, according to the positivists, is that the assumption is false. Not only do folk ordinarily engaged in moral practice presuppose the existence of moral facts, there are none. To be clear, the positivists did not thereby commit themselves nihilism. Their view was that when we make fundamental moral judgements, as we should, we express emotions in much the same way as we do when we say 'Boo!' or 'Hooray!' According to their *emotivism*, the

appropriateness of our doing this doesn't require that what we say is truth-apt. To put their point in terms of belief and desire, which are the basic psychological states they assumed constitute our emotions, the emotivists thought that when we make fundamental moral judgements, we don't express beliefs about the way the world is, but rather desires about the way it is to be.

The distinction just introduced between fundamental and non-fundamental moral judgements is important. This distinction tracks the familiar distinction between intrinsic desires, which are desires for states of affairs that are independent of any beliefs we have about natural facts, and extrinsic desires, which are desires that are dependent on such beliefs, as they are constituted by composites of intrinsic desires and beliefs. These could be beliefs to the effect that that the world's being a certain way would make it some way we intrinsically desire it to be, or beliefs to the effect that the world's being a certain way is a part of some whole way that we intrinsically desire it to be. For ease of exposition, let's call both these *means-end* beliefs. In terms of this distinction between intrinsic and extrinsic desires, the emotivists' view was that moral judgements could be expressions of either intrinsic desires (fundamental moral judgements) or extrinsic desires (non-fundamental moral judgements), and that natural facts, and our beliefs about them, though highly relevant to the appropriateness of non-fundamental moral judgements, are irrelevant to fundamental ones.

Suppose we believe that people governing themselves democratically would cause a state of affairs that we desire intrinsically, perhaps the state of affairs of there being more in the way of pleasure and less in the way of pain. In that case we might make either the fundamental moral judgement that more in the way of pleasure and less in the way of pain is better than the alternative, or the non-fundamental moral judgement that people governing themselves democratically is better than the alternative. According to the emotivists, the former expresses an intrinsic desire, the latter an extrinsic desire: that is, a combination of intrinsic desire and means-end belief. When we make the non-fundamental moral judgement, someone may therefore reject our judgement because they think the means-end belief we express is false. There is therefore important argumentative work for us to do in establishing that the natural world is a certain way to get those who reject our non-fundamental moral judgements to change their minds. In this case, for example, we would need to convince them that people governing themselves democratically causes more in the way of pleasure and less in the way of pain than alternatives.

This emotivist account of the role of argument in resolving moral disagreement was, however, strictly limited to resolving disagreements between parties who agree about fundamental moral judgements. In such cases one party may be able to get the other to change its mind by convincing them that their belief about what causes what, or what's a part of what, is in error. But given that fundamental moral judgements express intrinsic desires, the emotivists, who agreed with Hume that intrinsic desires are neither rational nor irrational, thought that all disagreement about them could amount to practical opposition. What's required in such cases is not a change of mind, as there is nothing for opponents to change their minds about, but rather a change of heart: a change in intrinsic desires. But if rational argument has come to an end in such cases, and we are left with the aim of changing hearts, then it seems that non-rational persuasion will be the only means available to us. In such cases, emotivism was committed to moral exchange being a matter of *goading*, not *guiding* (Falk 1953).

This idea, an idea that seemed to be the inevitable upshot of emotivism, came to seem incredible at the end of WWII (see also Lipscomb 2021). When the full scale of Nazi atrocities and the propaganda machine that they had mobilized became clear, it seemed too concessive to suppose

that a disagreement about fundamental moral principles like that between the Allies and the Nazis amounted to no more than a difference in their intrinsic desires, a difference to which arguments about truth and falsehood were irrelevant, but various modes of non-rational persuasion were highly relevant. By the late 1940s and early 1950s, many philosophers were therefore eager to find an alternative to emotivism, an alternative that allowed for more in the way of rational argument about fundamental moral judgements. The question was whether that would require a return to Moore's non-natural realism.

2. Normative ethics post-WWII: Rawls vs Hare

Among the philosophers who sought an alternative to emotivism were two young soldiers who returned post-war to the universities where they had completed their undergraduate degrees. As we will see, the work they published soon after, and the very different ways in which their work departed from emotivism, helped set the agenda for much of the work that has been done in normative ethics ever since.

On his return to the USA, John Rawls entered the PhD program in philosophy at Princeton where he went on to become an assistant professor. Soon after his return to England at much the same time, R. M. Hare was elected fellow and tutor in philosophy at Balliol College, Oxford. To anticipate, Rawls's response to emotivism was to turn his back on metaethics altogether by following *the method of avoidance*. He argued that the rational justification of our moral beliefs requires us to make no metaethical assumptions whatsoever. Hare's response was different. According to his *universal prescriptivism*, the emotivists were right that fundamental moral judgements are neither true nor false, and also right that intrinsic desires, not being dependent on beliefs, are neither rational nor irrational. But they were wrong that it follows from this that rational argument plays no role in resolving disagreements between those who accept different fundamental moral principles. Rawls and Hare both went on to make lasting contributions in normative ethics.

Let's begin with Rawls. Suppose we make some first-order moral judgement about a specific case, whether actual or hypothetical. How do we rationally justify the belief that that first-order moral judgement expresses? Rawls's answer in 'Outline of a Decision-Procedure in Ethics' (1951) was that, once we have settled the non-moral facts of the case, we ask ourselves which fundamental moral principle supports the judgement that we're disposed to make, and we then test the plausibility of this principle by seeing whether it delivers the correct verdicts—that is, the other first-order moral judgements we are disposed to make—about a range of other more specific hypothetical cases about which we are confident. In 'Outline', Rawls calls our judgements about such specific cases "considered judgements," and, as I understand it, his account of what makes judgements considered in 'Outline' is his attempt to spell out what makes such a judgement one in which our confidence is prima facie reasonable (more on this presently).

If the proposed principle delivers the correct verdict about these other cases, then we move on. But if it doesn't, then we ask ourselves whether we should amend the proposed principle so that it squares with our initial judgement about the specific case, or instead change our mind about which verdict is correct in the specific case. Again, as I understand it, Rawls's view is that this too is determined by our relative levels of confidence. Are we more confident about the proposed moral principle or about the verdict in the specific case? In 'Outline' Rawls suggests that we continue this process, sometimes revising our principles and sometimes revising our judgments

about specific cases, until we arrive at a relatively stable stopping point that he called a *narrow* reflective equilibrium.

Recognizing that a narrow reflective equilibrium might be dependent on our ignorance of alternative ways of thinking about matters, Rawls further suggested that we can and should consider the views of others in our community about which fundamental moral principles are correct, the verdicts that these principles deliver about specific cases, and the reasons that they give for these principles and verdicts, and that we should bring all of these into coherence with our own views as well. Since this could occasion further revisions of the fundamental moral principles about which we are most confident and their application to specific cases, Rawls called this a state of *wide* reflective equilibrium (Rawls 1974-75). In these terms, Rawls's answer to our initial question about rational justification was that, assuming knowledge of the non-moral facts of the case, the fact that a first-order moral belief about that case survives in a state of wide reflective equilibrium suffices for its rational justification.

On the face of it, this view of moral judgement rules out emotivism. As Rawls sees things, moral judgements about specific cases are derivable from judgements about fundamental moral principles together with judgements about the non-moral facts of the case, and judgements about fundamental moral principles can be brought into coherence with each another. Given that what makes such derivation and coherence relations possible is the fact that the contents of the judgements that stand in these relations are either true or false, and hence express beliefs, it follows that Rawls is committed to the truth of the very claim emotivists deny. It might therefore be thought that Rawls would get caught up in Moore's dialectic. Since his view is that there are moral beliefs and some are true, isn't he committed to the existence of moral facts, and since the only alternatives are that these facts are naturalistic or non-naturalistic, isn't he committed to taking a stand on which of these they are? But Rawls was later explicit that he isn't caught up in this dialectic (Rawls 1974-75). The rational justification of fundamental moral principles can and should, he tells us, proceed via *the method of avoidance*.

According to the method of avoidance, when we attempt to justify our moral judgements, we can and should, if possible, avoid taking a stand on any metaphysical issues. Since metaethics is a branch of metaphysics—Moore's non-naturalism is a theory about the nature of moral facts, and emotivism is a theory about the nature of the psychological states we express when we make moral judgements—Rawls's view was therefore that we can and should avoid taking a stand on whether moral judgements presuppose that moral facts exist, and what their nature is if they do, if we can. The payoff, according to Rawls, is that we will discover that we can rationally justify our moral judgements without ever taking a stand on these issues. 'But,' I hear you say, 'if Rawls is officially taking no stand on metaethical issues, what are we to make of the apparent tension between what he says about moral judgements when he gives his account their rational justification and what the emotivists say?' The answer turns on a distinction and an observation.

The distinction is that between folk-talk and philosophy-talk. What Rawls goes in for when he gives his account of the rational justification of moral judgement is folk-talk. The folk make first-order moral claims, they say that what they say is either true or false, they say that what they say expresses their moral beliefs, and they say that these beliefs can be either rationally justified or unjustified. Rawls's account of what it is for moral beliefs to be rationally justified is meant to be something that the folk can recognize and agree with. Of course, such folk-talk is either consistent with the philosophy-talk the emotivists go in for when they say that first-order moral claims don't express beliefs, and so aren't either true or false, or it isn't. But when someone takes

a stand on that issue of consistency they engage in philosophy-talk, and this is something Rawls tells us to avoid doing.

To give just one example, Simon Blackburn's preferred *expressivist* metaethical theory, which is in turn a sophisticated development of emotivism, has a component that he calls *quasi-realism* (Blackburn 1993). Quasi-realism purports to explain why, even though moral judgements are neither true nor false, and, at least in the case of fundamental moral judgements, express desires rather than beliefs, we would expect the folk to *say* that they express beliefs and are either true or false. Moreover, quasi-realism purports to explain why it would be okay for the folk to continue doing this. This is philosophy-talk par excellence. But Rawls's method of avoidance tells us not go in for any philosophy-talk about the nature of moral judgement, including philosophy-talk of this quasi-realist kind.

This is where the observation becomes important. Following the method of avoidance means that, even if we get our moral beliefs into a state of wide reflective equilibrium, there may well still be a gap in our beliefs—a failure of overall psychological coherence—that metaethical inquiry would be required to fill. More justificatory work, work of the kind that in their very different ways Moore and Blackburn try to do, therefore needs to be done. But it doesn't follow that that work needs to be done to rationally justify our first-order moral beliefs. The claims that we make when we go in for folk-talk and philosophy-talk are all claims about which we can be confident to a greater or a lesser degree. What Rawls claims to show with his account of wide reflective equilibrium is that we can leverage the coherence relations among the full set of first-order moral claims about which we are confident to get ourselves to quite reasonably change our minds, decreasing our confidence in some and increasing our confidence in others, and that when we engage in this process the metaethical claims we accept are irrelevant. They are irrelevant because, in Rawls's view, they aren't claims about which it is reasonable for us to be sufficiently confident for them to have any rational bearing on the—and here he speaks with the folk—*truth* of the moral claims we accept, given how confident we are about them.

To put the same point slightly differently—and here we speak with the folk on Rawls's behalf—since the moral beliefs that survive in wide reflective equilibrium would remain a part of our psychological profile in that same equilibrium even if we were to do the further justificatory work required to achieve overall psychological coherence, we can safely leave that further justificatory work to others. The method of avoidance codifies this consequence of the distinction between folk-talk and philosophy-talk and our relative levels of confidence in claims of both kinds. The upshot, if Rawls is right, is that once the non-moral facts of a case are fixed, the only evidence relevant to the rational justification of our moral beliefs is evidence of a first-order moral nature and evidence of our own reasonableness.

Given that so many normative ethicists ignore all metaethical considerations when they do normative ethics, and proceed via the method of reflective equilibrium, it should be clear why I said that much of the work done in normative ethics owes its character to Rawls's influence. Those who do work of this kind begin with their intuitions about the moral verdicts we should come to in various specific cases—these are the verdicts about which they are prima facie reasonably confident—and they use these verdicts to provide support for other first-order moral conclusions. These other conclusions might be conclusions about which fundamental moral principles are correct, or they might be conclusions about the verdicts in other specific cases about which they have, for the time being, suspended judgement. Thomas Nagel, Tim Scanlon, Michael Stocker, and Elizabeth Anderson are all notable examples of normative ethicists who do

work of this kind, and, as it happens, they were all Rawls's students. But other well-known figures who weren't students of Rawls also do work of this kind: Judith Jarvis Thomson, Derek Parfit, Frances Kamm, Shelly Kagan, Jefferson McMahan, Brad Hooker, Tim Mulgan, Seana Shiffrin, and Elizabeth Harman, to name just a view. Note that though some of these normative ethicists are deontologists, others are consequentialists. The methods of reflective equilibrium and avoidance do not themselves favour one substantive normative ethical theory over another.

Work in normative ethics of this reflective equilibrium kind is, however, not without its problems. Since such work begins with the moral judgements about specific cases, whether actual or hypothetical, about which theorists are confident—these reflect their *intuitions*—and then relies on coherence to force changes, there are two salient kinds of disagreement among theorists. The first is a disagreement between theorists who begin with the same intuitions, but who have different views about what coherence requires of them. Since at most one of them is right, we can readily understand why at most one of them is rationally justified and the other is not. This is not problematic. The second is a disagreement between theorists who begin with different intuitions, but then agree about what coherence requires of them. Since wide reflective equilibrium is silent about which intuitions we are to begin with, different theorists may be confident about different specific moral judgements at the outset, and these confidence levels may make a difference to the fundamental moral beliefs they wind up with in wide reflective equilibrium. Rawls's methods imply that both are rationally justified. This is problematic, as there is nothing to rule out someone's being confident about moral judgements that would ordinarily qualify them as being beyond the pale, and there is nothing to ensure that these aren't among those that survive in wide reflective equilibrium (Kelly and McGrath 2010).

Another problem with work of this reflective equilibrium kind concerns the levels of confidence Rawls claims it is reasonable for us to have about metaphysical, and hence metaethical, claims. Perhaps Rawls wasn't confident about any metaphysical claims, and perhaps none of those who follow him are confident either, but that is at best an idiosyncratic fact about them. Other theorists may well be confident, and there is no reason to think this makes them unreasonable. It is therefore simply false that the fact that a theorist's moral beliefs would survive in wide reflective equilibrium is sufficient for their rational justification. If the metaphysical beliefs of a theorist cannot be squared with their first-order normative beliefs, then the method of avoidance amounts to the injunction that they learn to live with an incoherent psychology, an injunction they would be irrational to obey if metaphysical arguments push them convincingly to revise.

Let's now turn to Hare's *The Language of Morals* (1952). Hare thought that emotivists were right that fundamental moral claims express desires rather than beliefs, and thus right that they are neither true nor false. He agreed that we don't ordinarily think of them in this way, but he thought that this is because we mischaracterize their semantic category. We think of claims like 'Killing John is wrong' as indicatives, when they are disguised imperatives: 'Killing John is wrong' is, he thought, another way of saying 'Do not kill John!' However, in addition to their being imperatives, Hare thought that moral claims have two further distinguishing marks. This is where he went beyond emotivism. One is that those who make moral claims about specific cases commit themselves to a further universal moral claim, a claim that applies not just to the specific case, but to all cases with the same universal non-moral features as the specific case. The other is that those who make moral claims issue imperatives that are overriding. According to Hare, it is because moral claims have these further distinguishing marks that we can rationally come to change our minds about which fundamental moral claims we accept.

Suppose I say, 'Killing John is wrong'. Hare thought that I thereby commit myself to making the same claim about—that is, expressing the same overriding desire about—all cases that have the same universal non-moral features as the specific case of killing John. If I think that the relevant universal feature of the specific case that makes it the case that killing John is wrong is that John is an innocent person, then the universal claim in the background would be that it is wrong to kill innocent people, and this would have to express an overriding desire I have that innocent people are never killed. But, as Hare points out, once I vividly imagine all of the scenarios in which an innocent person is killed, cycling through what it would be like for each of the people who exist in those scenarios, and compare my reactions to these to my reactions to the scenarios that would have obtained if that innocent person had not been killed, cycling myself through what it would be like for each of the people who exist in those scenarios, then I might well find that I would prefer some of the latter scenarios to the former. (Imagine that if the innocent person is not killed then a million people will be tortured for the rest of their lives.) If so, then coherence would require me to get rid of my overriding desire that innocent people are never killed. This would be to rationally change my mind about one of my fundamental moral judgements.

Hare was a utilitarian in normative ethics. Getting someone to imagine what it would be like for all those affected by the different actions available to an agent in a scenario with certain universal non-moral features would therefore, he thought crucially, get them to sympathetically identify with those affected. The result of this sympathetic identification would, he thought, be the formation of a preference that the action that best satisfies the weighted preferences of all those affected be performed in all of the scenarios identical in their universal non-moral features. Hare was no doubt right that some people who were not antecedently utilitarians would come to be utilitarians by doing such sympathetic imagining, but he recognized that some would not, and in the early part of his career he sorted these people into two classes. In one class are those who form a preference that some other action be performed in all such scenarios. Hare labelled these people *fanatics*. In the other class are those who wind up having no intrinsic desire that an action be performed in all scenarios identical in their universal non-moral features. Hare labelled these people *amoralists*.

It is clear why Hare called those without any intrinsic desires with universal contents "amoralists." To make moral judgements, he thought, is to express universal desires, so those who have no universal desires make no moral judgements at all. Moreover, since Hare thought that intrinsic desires are neither rational nor irrational, he thought that no argument could get an amoralist to acquire any intrinsic desires with universal contents. They were therefore beyond the moral pale. "Fanatic" is, however, a pejorative term, and it was unclear why Hare felt entitled to use it of those who had intrinsic desires with universal contents, where those contents were not utilitarian in character. Being a utilitarian himself, he of course disapproved of those with such universal desires. But that hardly suffices for labeling them fanatics.

Having said that, it is noteworthy that later in his career Hare appealed to a principle that he claimed to be a conceptual truth—*If someone actually knows that he would prefer that p with strength s if he were in circumstances C, then he actually prefers that p in C with strength s*—to argue that the only thing anyone can consistently desire be done in all scenarios identical in their universal non-moral features, at least once they have gone through the imaginative exercise he describes, is the action recommended by utilitarianism (Hare 1981). He therefore came to think that those he had earlier labelled fanatics suffered from either a failure to properly imagine all the scenarios about which they had a universal desire, or an inconsistent overall psychology: they

were either ignorant or irrational. However few thought that Hare was right about this. The principle on which he relied plainly wasn't a conceptual truth, and if he retreated to the weaker claim that it expresses a normative principle—a principle of rationality, perhaps—then he would have had to take issue with the Humean view that intrinsic desires are neither rational nor irrational. This was something he was disinclined to do given that he thought those who have no desires with universal contents at all are amoral, but not irrational.

Another view Hare argued for later in his career was, however, quite influential. Hare thought that there is both critical level moral thinking, which is done in a cool hour, free from constraints of time and energy, and intuitive level moral thinking, which is the kind of thinking that we engage in when we make moral decisions day to day, and so is constrained by time and energy. In these terms, Hare thought that critical level moral thinking is what we're doing when we test our universal desires for consistency having gone through the imaginative exercise described earlier, and that the precise nature of the intuitive level moral thinking that we should engage in would therefore depend on what we learn when we do our critical level moral thinking. What kind of day-to-day decision-making procedures are supported by the principles we accept after having engaged in critical level moral thinking? Since he thought that we should accept utilitarianism at the critical level, his distinction between levels of moral thinking was a version of the familiar distinction utilitarians since Mill had made between a *criterion of rightness* and a *decision-procedure* (Mill 1863). Hare thought, much as Mill and others had thought, that utilitarianism is in the first instance the former, and the latter only when time and energy allow.

Note how very different Hare's view is from Rawls's. Hare begins with the metaethical view that moral claims are universal overriding imperatives, and he then argues (unconvincingly, but still) that if we make any moral judgements at all, the only ones we can consistently make are those that accord with utilitarianism. Importantly, someone could rationally come to this utilitarian conclusion on the basis of Hare-style sympathetic identification without ever having had antecedent utilitarian intrinsic desires. The contrast with emotivism was thus significant indeed. But this kind of metaethical argument for a normative ethical conclusion would have been unavailable to Hare if he had followed Rawls's methods of reflective equilibrium and avoidance. Indeed, Hare's argument, if it were convincing, would justify outright hostility towards Rawls's methods (Hare 1972). According to Hare, our antecedent moral convictions constitute intuitive level moral thinking. As such, they may be fine for making moral decisions under pressure of time or in situations of low energy, but it all depends on whether critical level moral thinking supports their use in such situations.

I said at the outset that Hare's influence also explains why much of the work done in normative ethics has the character that it has. Hare's most famous student was Peter Singer whose early work in applied ethics was inspired by Hare's own (Singer 1979). Given the influence of Singer's work on those who work in normative ethics, and indeed on the moral beliefs of legions of ordinary folk, it is therefore hard to overstate the impact, albeit indirect, of Hare's views. The main feature of all such work is its zeal for utilitarianism. There is an emphasis on moral thinking done in a cool hour when our sympathetic imaginations can be engaged, coupled with a steely refusal to shy away from the controversial conclusions to which such thinking may lead. Objections based on antecedent moral convictions are dealt with by arguing that such convictions may have a role to play when there are constraints of time and energy, but only to the extent that acting in accordance with them is utility maximizing. When it is not, acting in accordance with our antecedent moral convictions lacks a critical level rationale.

This kind of work in normative ethics is, however, not without its problems either. From the perspective of the one who makes moral claims, the main problem arises because of the often implicit assumption, an assumption about which Hare himself was quite explicit, that moral judgements are overriding. When you couple this with utilitarianism, the upshot is that it is okay for you to do what we ordinarily think of as something that it is your prerogative to do—to form relationships with friends and lovers, to produce or consume art, to spend time enjoying nature, and so on—only if either these activities are themselves utility maximizing, or it is utility maximizing for people to act as if they had a prerogative over such matters. Since neither of these conditions would typically be met, this turns into the well-known, and for many decisive, demandingness objection to utilitarianism.

From the perspective of those about whom one makes moral claims, the problem is different. Once you allow your metaethical views to drive your approach to normative ethics, there can be no picking and choosing among the metaethical views that do the driving. The issue on which Hare took a stand was one that had been front and center since Moore and the emotivists. Does moral judgement presuppose the existence of a domain of moral facts? Hare agreed with the emotivists, and against Moore, that they do not. But another metaethical issue on which we must take a stand is *moral rationalism*—that is, whether moral requirements entail normative reasons for action (in what follows I will take the qualifier 'normative' as read)—and the Humeanism that lies in the background of Hare's view commits him, as it committed the emotivists, to moral anti-rationalism. This is problematic, as the claim that an act is morally wrong but there is no reason not to do it sounds self-contradictory. Suppose someone fails to maximize utility because they spend all of their time producing art. Though Hare thinks that they have acted wrongly, his moral anti-rationalism commits him to the conclusion that they may have done something that they had every reason to do and no reason not to do. It all depends on what their intrinsic desires are.

Let's sum up. The discussion of Rawls and Hare prompts the following important questions. Suppose we agree with Moore and Hare about the significance of metaethics to normative ethics—that is, suppose we reject Rawls's method of avoidance—but we start from a very different metaethical premise from them, namely, moral rationalism. Will we end up thinking, with Hare, that there are no moral facts, or with Moore that there are and that they are non-natural? Will we end up thinking, with both Moore and Hare, that utilitarianism is the correct normative ethic? And finally, will we end up accepting or rejecting Rawls's account of the rational justification of our moral beliefs in terms of reflective equilibrium?

3. From moral rationalism to the ethics of agency

Moral rationalism is the view that moral requirements entail corresponding reasons for action. The immediate questions to which we require answers are the following. Is moral rationalism defensible? If so, does this entail that there are moral facts? If so, are the entailed moral facts naturalistic? And finally, do any first-order moral claims follow from all of this? Recent work on *constitutivism* suggests positive answers to these questions (Smith 2018).

Let's begin from the best-known version of the opposing anti-rationalist view of reasons for action: Bernard Williams's Humean account (1981). Williams begins from the plausible idea that what agents have reason to do is fixed by what they would want themselves to do if they weren’t vulnerable to reasoned criticism, and then adds a distinctively Humean twist. Simplifying somewhat, he tells us that reasoned criticism can only amount to either the criticism of agents' beliefs as false or epistemically unjustified, or the criticism of their extrinsic desires for failing to

reflect their intrinsic desires and their means-end beliefs (remember the earlier stipulation that means-end relations are to be understood as including beliefs about both causal and part-whole relations). This reflects Williams's commitment to the Humean view that intrinsic desires are neither rational nor irrational. The upshot is that what agents have a reason to do is whatever they want themselves to do on condition that their wants aren't based on ignorance or error or means-end irrationality, and this will inevitably reflect the intrinsic desires with which they began. If we can explain epistemic reasons in truth-based terms, then Williams's Humean account of reasons for action entails both that there are facts about what agents have reasons to do and that these are natural facts.

There is, however, a problem we can bring into focus if we restate the Humean view of reasons in terms of the implicit account of an ideal agent on which it relies. This is where recent work on constitutivism becomes relevant. According to the Humean, there are two kinds of agential capacity. When we criticize agents for being ignorant or in error, we criticize them for failing to exercise their knowledge-acquisition capacities, and when we criticize them for being instrumentally irrational, we criticize them for failing to exercise their capacity for intrinsic desire-realization: that is, their failure to put their intrinsic desires together with their means-ends beliefs. This means that we can define an agent as someone who possesses these two agential capacities, and that we can further define an ideal agent—that is, someone who meets all of the standards of assessment that are internal to being an agent—as someone who robustly possesses and exercises maximal agential capacities. An agent's capacity to know what the world is like is maximal when it isn't contingent on the world's being certain ways rather than others—that is, they have the capacity to know what the world is like, no matter what it is like, to the extent that it is knowable—and an agent's capacity to realize their desires is maximal when it isn't contingent on their having intrinsic desires for certain things rather than others, or certain abilities rather than others— that is, they have the capacity to realize their intrinsic desires in the world, no matter what their content, to the extent that their intrinsic desires are realizable.

The Humean account of reasons for action can now be restated in terms of the desires of the ideal agent, so understood. The attraction of restating the Humean view in this way should be clear. Reasons for action are normative, but what is the source of their normativity? The restatement provides an answer. Reasons for action inherit whatever normativity they have from the normativity that is internal to the concept of an agent: that is, they inherit their normativity from the *ideality* of ideal agents. Ideal agents have plans for every contingency, including the contingency in which they have non-maximal agential capacities. What non-ideal agents in certain circumstances have a reason to do is therefore to put their ideal agential counterparts' plans for what to do those circumstances into action. If a non-ideal agent's ideal counterpart's plans reflect the intrinsic desires of the non-ideal agent, then this is equivalent to Williams's Humean view of reasons for action. However, as we will see, this Humean conception of the ideal agent entails a contradiction (Smith 2020).

On the one hand, Humeans think that agents' intrinsic desires are neither rational nor irrational. Whether an agent is ideal is thus supposed to be independent of their having or lacking certain intrinsic desires. Non-ideal agents can have any old intrinsic desires, and their ideal counterparts will have all and only those same intrinsic desires. But on the other hand, Humeans are also committed to the view that what agents have reasons to do is fixed by what their ideal agential counterparts want them to do, where their ideal agential counterparts robustly have and exercise maximal capacities for knowledge-acquisition and desire-realization. Ideal agents meet this

robustness condition when a whole range of non-ideal worlds that are nearby the ideal world—that is, a whole range of worlds in which agents are only a little bit non-ideal, and a little less ideal than that, and so on—are worlds in which they still score highly on the scale of knowledge-acquisition and desire-realization. The problem with the Humean conception of an ideal agent arises because, as we will see, meeting this robustness condition requires that ideal agents have certain intrinsic desires. The Humean view thus entails a contradiction. Ideal agents can have any old intrinsic desires, but they must also have certain intrinsic desires.

The argument that meeting the robustness condition requires ideal agents to have certain intrinsic desires, and that the worlds in which they find themselves must meet certain other conditions, proceeds as follows. Since agents are temporally extended, those who exercise their agential capacities robustly must have and exercise these capacities at each moment they exist. Agents who exercise their agential capacities robustly must therefore have the wherewithal within themselves, or the worlds in which they exist must contain the wherewithal, to overcome two kinds vulnerability to which they would otherwise be subject, vulnerabilities that would otherwise prevent them from having and exercising their agential capacities at each moment they exist.

The first kind of vulnerability is the vulnerability of an agent's later self to their earlier self. Imagine an agent at a certain time who exercises the capacity to know what the world is like and satisfy their desires at that time. For their exercise of their agential capacities to be robust, it cannot be the case that that exercise is dependent on the fact that, at an earlier time, they just so happened not to want to deceive themselves at the later time, or that at the earlier time, they just so happened not to want to undermine the satisfaction of their desires at the later time. It cannot because this would make their possession and exercise of their agential capacities fragile: that is, it would entail that in worlds nearby the ideal world in which they lack those desires, they don't exercise their agential capacities. Moreover, if their possession of their agential capacities is itself the product of their efforts to develop and maintain them at some earlier time, then it couldn't be dependent on the fact that, at that earlier time, they just so happened to want that they develop and maintain such capacities that they could exercise at the later time. This too would make their possession and exercise of their agential capacities fragile. They must instead have some feature, in virtue of being an agent who robustly has and exercises maximal agential capacities at each moment they exist, that explains why at the earlier time they helped develop and maintain their having their agential capacities, and why at the earlier time they didn't interfere with their later exercise. But what feature of an agent could possibly explain this?

The most straightforward feature to imagine them having at the earlier time is a pair of intrinsic desires that are constitutive of being an agent whose possession and exercise of their agential capacities is maximal and robust. They must have an intrinsic desire that they have agential capacities to exercise, and they must also have an intrinsic desire not to interfere with their exercise of their agential capacities once they have them on condition that that exercise won't itself constitute interference (from here-on I will take this condition as read). Another obvious feature to imagine them having is the self-control required to acquire and realize such intrinsic desires in nearby worlds in which they are subject to apathy or weakness of will. We could then suppose that agents who lack these two intrinsic desires, or who lack the capacity for self-control, are subject to reasoned criticism simply on that account. They are subject to reasoned criticism because, absent their possession of these intrinsic desires and the capacity for self-control, their possession and exercise of their agential capacities would not be maximal and

robust. But note that to say this is to go beyond the Humean account of what reasoned criticism of an agent can amount to. Certain intrinsic desires will be required to avoid reasoned criticism.

The second kind of vulnerability they must have is the wherewithal to overcome, or the world in which they exist must contain the wherewithal for them to overcome, the interpersonal analogue of the intrapersonal vulnerability just described. Imagine an agent at a certain time who exercises the capacity to know what the world is like and satisfy their desires in it at that time. For their exercise of this capacity to be maximal and robust, it cannot be the case that that exercise is dependent on the fact that, at that time, there are other agents who just so happen not to want to deceive them, or that at that time, there are other agents who just so happen not to want to undermine their satisfaction of their desires. That would make their exercise of these capacities fragile. Moreover, if their possession of those capacities at that time is the product of the efforts of other agents to help them develop and maintain their capacities, it couldn't be dependent on the fact that those other agents just so happened to exist and want to help them develop and maintain their capacities. Those other agents must instead have some further feature that explains why they exist and help but don't interfere.

This suggests that we would be wrong to suppose that agents could robustly possess and exercise maximal agential capacities in worlds in which there are no other agents, or in which there are other agents but they have just any old intrinsic desires. Moreover, it also suggests that we were wrong when we earlier supposed that the intrinsic desires that agents must have to overcome their intrapersonal vulnerabilities were desires concerning just their own agential capacities. For any particular agent's exercise of their agential capacities to be robust, they must be in the company of other agents who robustly possess and exercise maximal agential capacities, and they must all intrinsically desire that every agent has agential capacities to exercise, and also intrinsically desire not to interfere with the exercise of the agential capacities of any agent. Moreover, all such agents must have the capacity for self-control required to ensure that, in nearby worlds in which they are subject to apathy or weakness of will, they acquire and realize such intrinsic desires anyway. In sum, to be an ideal agent is to be in the company of other similarly ideal agents all of whom desire that people have agential powers to exercise, and all of whom desire that they themselves do not interfere with their exercise. This explains why ideal agents are as invulnerable as any finite being could be to interference and loss of their agential powers. In other words, this explains why ideal agents *robustly* possess and exercise maximal agential capacities.

Let's take stock. If you begin with the Humean anti-rationalist idea that what agents have reason to do is fixed by what they would want themselves to be doing if they acted in ways that avoid reasoned criticism, where an agent's intrinsic desires, being neither rational nor irrational, are beyond reasoned criticism, then you will inevitably be led to reject that view in favour of an anti-Humean rationalist view. All agents do indeed have reasons to do what they would want themselves to be doing if they acted in ways that avoid reasoned criticism, and this entails that they have reasons to do what their ideal counterparts would want them to do in their circumstances, where their ideal counterparts' being ideal is a matter of their robustly possessing and exercising maximal agential capacities. But this in turn entails that they have substantive reasons to do three things: they have reasons to help make sure that everyone has agential capacities to exercise; they have reasons not to interfere with anyone's exercise of their agential capacities once they have them; and, beyond that, they have reasons to satisfy whatever intrinsic desires they have.

If we suppose that Kant is right that the mark of a *moral* reason for action is that it is unconditional (Paton 1948)—that is, that it is a reason to do something in certain circumstances whether or not you have a desire to act in that way in those circumstances—then it turns out that everyone has two moral reasons for action, namely reasons (for short) to help but not interfere, and that they also have non-moral reasons to do whatever they want to do. This anti-Humean rationalist view of what agents have reasons to do should sound familiar, as it is basically the Enlightenment idea that Kant himself defended. According to that Enlightenment idea, everyone has a reason to live a life of their own choosing unencumbered by others, a reason to ensure that everyone has the wherewithal to make such life-choices, and a reason, once they have that wherewithal, not to interfere with their acting on whatever life-choices they make. There are, however, some important differences from Kant's view that need to be kept in mind.

For one thing, though the class of beings who have reasons to help but not interfere—that is, the class of moral *agents*—is restricted to the class of agents who are conceptually sophisticated enough to think of themselves as having the options to help and not interfere, the class of beings who these agents have reasons to help but not interfere with—that is, the class of moral *patients*—is the entire class of agents. Everyone with the capacity to know what the world is like and to realize their desires in it—mature humans, infant humans, humans with serious cognitive impairments, and all manner of non-human animals—falls within the scope of moral reasons. The contrast with Kant's view that moral agents and patients must both have the capacity to act on the categorical imperative could not be more stark. There is another difference from Kant's view as well. Kant was famously hostile to a teleological conception of reasons: that is, a conception of reasons as sourced in the value of outcomes. But given that the reason to help is explained by the ideal agent's desire that everyone has agential capacities to exercise, it follows that that reason must be understood in teleological terms. There is agent-neutral value in beings having agential capacities to exercise—that follows analytically from the ideal agent's having the corresponding agent-neutral desire—and the reason to help is a reason for action because it realizes that agent-neutral value. As we will see in a moment, this suggests that all three reasons are to be understood teleologically, and that a further difference between the two moral reasons therefore lies in the fact that, whereas the reason to help realizes an agent-neutral value, the reason not to interfere realizes agent-relative value (Smith 2009).

A number of further clarifications are in order. The first is that there is evidently indeterminacy in the content of the two moral reasons as specified so far. Consider the reason not to interfere with anyone's exercise of their agential capacities, by way of example—similar points could be made about the reason to help. If you cause someone to be in extreme agony, then they may be so overwhelmed that they are unable to act at all. But now imagine a continuous sliding scale of pains with extreme agony at one end and only the faintest feeling of pain at the other, a feeling so faint that it doesn't interfere with anyone's exercise of their agential capacities, and then imagine a range of actions that cause the pains that lie at each point along this sliding scale. At what point do these actions amount to interference with an agent's exercise of their capacity to know what the world is like and realize their desires in it? The answer is that there is no such point. There is instead a range of actions that are determinately instances of interference, and a range of actions that are determinately not instances of interference, and in between there is a range of cases about which it is indeterminate whether they are cases of interference.

The second clarification concerns the weights of the three reasons as so far specified. To stick with the example of the reason not to interfere—similar points could be made about the other

two reasons—this reason will be more or less weighty depending on the extent to which the interference undermines an agent's exercise of their agential capacities. Given that the three reasons can conflict with each other, this means that there is no easy way to spell out the relative weights of the three reasons in cases where they conflict. To return to the Enlightenment idea, the best thing we can say is that the option best supported by all three reasons taken together will be that option in which the mix of living a life of one's own choosing, not interfering with anyone's leading a life of their own choosing, and making sure that everyone has the wherewithal to lead a life of their own choosing, is that which is most similar to the mix in the world in which all agents are ideal. In some cases, the moral reason to help will outweigh the moral reason not to interfere, in other cases vice versa. In some cases, the moral reasons to help but not interfere will outweigh the non-moral reason to lead a life of one's own choosing, and in other cases vice versa. Moreover, since there is bound to be vagueness in this similarity relation, the weights of these reasons will almost certainly be vague too.

It is this feature of the three reasons— that they can be weighed against each other and their weights can vary—that tells in favour of their being understood in teleological terms. So understood the three reasons for action are grounded in values realized by acting on them. The reason we have to help is grounded in an unconditional agent-neutral value: that is, a worldly state we intrinsically desire when we are ideal that requires no mention of us in characterizing it. The reason we have not to interfere is grounded in an unconditional agent-relative value: that is, a worldly state we intrinsically desire when we are ideal that does require that we be mentioned in characterizing it. And the reason we have to do whatever we want is grounded in a conditional value that is either agent-neutral or agent-relative, depending on the content of the other intrinsic desires that we just so happen to have when we are ideal. The benefit of thinking of the three reasons in these teleological terms is that the weights of the three reasons can be understood in terms of a combined ranking of the outcomes of acting on them in terms of the amount of value they realize (compare Jackson and Smith 2006). The view could also be stated in terms of principles. There are two pro tanto moral principles, one—*Everyone ought to help*—reminiscent of a consequentialist moral principle, and the other—*Everyone ought not to interfere*—reminiscent of a deontological moral principle, and a pro tanto non-moral principle—*Everyone ought to do whatever they want*—reminiscent of Humeanism about reasons for action, and what we ought to do all things considered is fixed by weighing these principles against each other in the circumstances, something that is only made possible by thinking of them all in teleological terms (compare Portmore 2011, and for criticisms that require a response see Kiesewetter 2022).

The third and final point that needs to be made about the three reasons for action is that, given what's just been said about their contents and weights, we can provide the following plausible definitions of the basic deontic statuses of actions—I will explain what non-basic deontic statuses are presently—that can be thought of as partially defining a new normative ethic that, for want of a better name, I will call the *ethics of agency*. Morally wrong actions are those there are decisive moral reasons not to perform, where reasons are decisive just in case they determine what we have all things considered reason to do; morally permissible actions are those there are no decisive moral reasons not to perform; and morally obligatory actions are those there are decisive moral reasons to perform.

We are now in a position to answer the questions with which we began this section:

Is moral rationalism defensible? Moral rationalism is the view that moral requirements entail corresponding reasons for action. Moral requirements are all statable in terms of the basic

deontic statuses—being morally wrong, permissible, and obligatory—and as we have seen, these are all definable in terms of the relative weights of moral and non-moral reasons for action. The entailment thus holds.

Does the defence of moral rationalism entail that there are moral facts? The defence of moral rationalism does entail that there are moral facts because it entails that there are facts about what agents have moral reasons to do.

Are moral facts naturalistic or non-naturalistic? Moral facts are naturalistic facts because facts about what agents have moral reasons to do turn out to be facts about the desires that all agents have simply in virtue of being ideal, where being ideal is itself naturalistically characterizable.

Do any first-order moral claims follow from the truth of moral rationalism? Moral rationalism entails the first-order ethics of agency because it entails that all agents have unconditional reasons to help but not interfere, and conditional reasons to do whatever they like. These are the core first-order claims of the ethics of agency from which further claims about which actions are morally wrong, permissible, and obligatory in various circumstances follow.

4. Additional normative ethical questions raised by the ethics of agency

As we have seen, the contents and weights of the reasons to help, but not interfere, and apart from that to do what we like, are vague. An important question for the ethics of agency is whether the indeterminacy of these reasons reflects the limits of what we can learn about the content and weight of our reasons on the basis of reflection, or is instead an artifact of the specific metaethical argument given for their existence. Let's consider these ideas in reverse order.

Though no role has so far been played by reflective equilibrium reasoning in arguing for the contents and weights of the three reasons, a residual role can and should be played, as reflective equilibrium reasoning enables us to draw on other considerations that may be relevant to fixing these contents and weights. The inputs to such reflective equilibrium reasoning would include the metaethical considerations we adduced in arguing for them, but it would also include all of the arguments that have been given for the different conceptions of Enlightenment values that the three reasons reflect—freedom, equality, fraternity, respect, dignity, autonomous choice, strength of will, self-control, and so on—as these play out in the various circumstances in which the three reasons must be weighed against each other. If we figure out which conceptions of these Enlightenment values are best integrated with the metaethical argument given for the three reasons, and the circumstances in which they conflict, then we may well discover that the reasons have much more determinate contents and weights than they seem to have based only on the metaethical argument. At the limit, we may discover that these contents and weights are fully determinate.

It is, however, important to keep in mind how very different this use of reflective equilibrium reasoning is from Rawls's. No role is played by the method of avoidance, and no role is played by moral judgements just because we or others have confidence in them. Indeed, the connection between the three reasons and Enlightenment values suggests a test for which of moral judgments we are confident about, independently of having any arguments for them, can reasonably serve as inputs to reflective equilibrium reasoning. If those antecedent moral judgements reflect some vague sense of Enlightenment values, then they can reasonably serve as input, but if they do not then they cannot. Note that this means that reflective equilibrium

reasoning of the kind recommended here does not face one of the problems that plagues Rawls's account. Considered judgements that reveal one to be beyond the pale cannot reasonably serve as an input to reflective equilibrium reasoning because such judgements don't reflect a vague sense of the Enlightenment values supported by the metaethical arguments. Note too that, though Hare turns out to be right that reasoning at the critical level is in the first instance metaethical, he is wrong that no role can or should be played by reflective equilibrium reasoning.

As I said, it is possible that reflective equilibrium reasoning will deliver fully determinate conceptions of the contents and weights of the three reasons, but it may not. This brings us to the other idea. The vagueness of the three reasons may reflect the limits of reflection for learning about the contents and weights of our reasons for action. If so, then that provides us with a practical problem, as we still need to act over stretches of time, and to coordinate with others, and acting in these ways without tripping over ourselves requires us to settle on more determinate contents and strengths. Two strategies suggest themselves. The first is personal commitment. Each person could commit themselves to acting as if the three reasons had a certain determinate content and weight. This may solve the diachronic problem, and if few people are involved and they all know what each other has personally committed to, then it might enable them all to coordinate their actions with each other too. But in circumstances in which many people are involved, and knowledge of each other's personal commitments is therefore impossible, the commitment strategy will not work.

What is needed in such circumstances is a set of social rules that fix a determinate conception of the contents and weights of the three reasons to be acted on in circumstances in which our actions affect others about whom we know very little (Smith 2021). Since social rules are kept in place by threats of sanction for non-compliance, we can specify a pair of conditions that need to be met by such non-compliant actions. First and foremost, they must be actions which are morally wrong in the basic sense defined earlier— that is, they must be actions that there is a decisive moral reason for people not to perform—and second, they must be actions that it is morally permissible in the basic sense for some yet-to-be-specified people to sanction wrongdoers for performing—that is, there must be no decisive moral reason for such people not to do so. Let's call actions that meet this pair of conditions *juridical moral wrongs*. Non-basic moral deontic statuses like juridical moral wrongness raise a host of further questions for ethicists of agency.

For one thing, we need to know what conditions must be met by the yet-to-be-specified people who are morally permitted to sanction juridical moral wrongdoers, and what conditions must be met by the sanctions. In ordinary parlance, this is a question about the *authority* of these people to impose those sanctions. What gives them that authority? Does that authority itself require the existence of social rules? For another, we also need to know how juridical moral wrongness is affected by the existence of social rules that get people to perform actions that do not meet these two conditions. What are we to do when there exist social rules, when acting in a way that is morally permissible requires the existence of social rules, but the social rules that exist aren't the ones that must exist for us to act morally permissibly? My hunch is that many of the moral wrongs discussed by normative ethicists are juridical moral wrongs of this kind. If so, this would explain why the views taken on such moral wrongs occasion hostile reactions from members of the general public who sincerely believe that it is juridically morally obligatory to comply with the existing social rules. It also raises a further question for ethicists of agency. What exactly

does the reason not to interfere tell in favour of in such circumstances? As I understand it, this is the vexed issue of the nature and value of *civility* (Calhoun 2000).

A final question to which we need an answer is which juridical moral wrongs it is not only morally permissible, but morally obligatory, for some yet-to-be-specified people to sanction juridical wrongdoers for performing. What is at issue here is the scale of moral wrongs. What determines the scale of moral wrongs such that that no-one has the authority to sanction moral wrongdoers for performing certain moral wrongs, other moral wrongs are such that some people have a moral permission but not a moral obligation to sanction moral wrongdoers for performing them, and yet other moral wrongs are such that some people have a moral obligation to sanction moral wrongdoers for performing them? Discussions of moral responsibility that assume a tight connection between moral wrongness and the appropriateness of sanctions—for example discussions inspired by P. F. Strawson's "Freedom and Resentment" (1962)—are, it seems to me, vitiated by a failure to distinguish clearly between moral wrongs of these three kinds.

Reflection on the two ways of making indeterminate moral reasons more determinate— the discovery of a principled precisification of their content and weight via reflective equilibrium reasoning and the arbitrary precisification of their content and weight via social rules—may illuminate some existing problems in normative ethics. To give one example, The Trolley Problem suggests that ordinary folk recognize that there are indeed the two moral reasons we have identified—an agent-neutral reason to help and an agent-relative reason not to interfere—and that they further recognize that in some cases (sacrifice the one to save the five) the agent-neutral reason outweighs the agent-relative reason, while in other cases (do not kill the one even though the five will die) the agent-relative reason outweighs the agent-neutral reason. But the difficulties involved in deciding what to think about some of the many variations on The Trolley Problem might show one of two very different things. One is how difficult it is to discover the principled precisifaction of the content and strength of the two reasons via reflective equilibrium reasoning. The other is that a principled precisification is unavailable, and that what's required in such cases is a social rule that currently doesn't exist.

So far we have focused on a range of first-order moral issues raised by the ethics of agency. These are all cases in which there are either decisive moral reasons for people to act in a certain way, or no decisive moral reason not to act in a certain way. But cases of the latter kind divide into sub-classes of which two are of special interest. In one, the moral reasons for acting in a certain way are on a par with the reasons for not acting in that way, where the reasons for not so-acting are all non-moral in nature. In the other, the moral reasons for acting in a certain way are outweighed by the reasons for not acting in that way, where the reasons for not so-acting are all non-moral in nature. In the latter sub-class, it is morally permissible to perform an action that there is a decisive non-moral reason not to perform. This in turn suggests is that the ethics of agency requires us to acknowledge the following neglected non-moral deontic statuses.

There are actions that are non-morally wrong: these are actions that there are decisive non-moral reasons not to perform. There are actions that are non-morally permissible: these are actions that there are no decisive non-moral reasons not to perform. And there are actions that are non-morally obligatory: these are actions that there are decisive non-moral reasons to perform. In these terms, the acts identified in the second of the two sub-classes of moral permissions identified above are morally permissible but non-morally wrong. Which actions have these non-moral deontic statuses is a question that normative ethicists can and should answer. This suggests how very different the agenda in normative ethics would be if it were being done by ethicists of

agency from the way normative ethics being done by utilitarians like Hare and Singer and those they have inspired. Think again about cases in which we are wondering whether to form relationships with friends and lovers, or whether to produce or consume art, or whether to spend time enjoying nature, but there are moral reasons not to do so on the side. In certain such cases, we may not just have a moral prerogative to act contrary to our moral reasons. It may be non-morally wrong for us not to do so (compare Wolf 1982).

The category of non-moral wrongs also promises to illuminate a scenario that isn't much discussed by philosophers, but that features heavily in science fiction. I will call this the *brutal apocalypse*. A brutal apocalypse is a scenario in which the vast majority of people meet four conditions: (i) they aren't reliably disposed to do what they have most reason to do, as they don't care about morality; (ii) their relationships are all transactional, so they have neither lovers nor friends; (iii) they have very little in the way of impulse control; and (iv) they lack the wherewithal to change any of these features of themselves via rational reflection. Note that a brutal apocalypse may or may not be a *state of nature*. A state of nature is a world in which there are no social rules, where social rules require the existence of a sub-group who makes, promulgates, and polices the rules (Smith 2021). Since the sub-groups needn't act in the ways they do for moral reasons or for reasons of love or friendship, and since they needn't have much in the way of impulse-control, and since a single powerful person could make the rules on a whim, and then terrorize another sub-group into promulgating and policing those rules on his behalf, it follows that some brutal apocalypses are not states of nature.

Now imagine that you are a member of the very small minority in a brutal apocalypse who doesn't meet the four conditions, and, just to keep things simple, imagine that the brutal apocalypse in which you find yourself is a state of nature. What would you have most reason to do? Asking this question focuses our attention on the ways in which the relative weights of our moral and non-moral reasons are affected by the fact that the moral patients of our acts do or don't care about other people, the fact that they do or don't have impulse control, and the fact that they lack the wherewithal to change these features of themselves via rational reflection. My hunch is that in such circumstances you may have no decisive moral reason not to search out others who are like you with whom you could form bonds of love and friendship, and that, once you have formed such bonds, you and your lovers and friends may have no decisive moral reason not to act on your own non-moral reasons in many situations, or on the moral reasons you have to help and not interfere with each other rather than the rest. Indeed, my hunch is that it may well be non-morally wrong for you not to act in these ways that favour each other in many situations. The point is not that you have no moral obligations towards others in the brutal apocalypse. The point is rather that you have far fewer than you would have if conditions (i)-(iv) weren't met.

If these hunches are correct, then there are important lessons to be learned. One lesson concerns the behaviour of those in the here-and-now who act as if we were in a brutal apocalypse. Since the here-and-now isn't a brutal apocalypse—the four conditions are not met by the vast majority—it follows the actions of such people are morally wrong. But having said this, a question to which we require an answer does concern the here-and-now. Since we can imagine a whole series of worlds, each only slightly dissimilar to those adjacent to it, that takes us all the way from the brutal apocalypse to the here-and-now, we must ask whether the respects of similarity between the brutal apocalypse and the here-and-now already justifies some weighing

in favour of those with whom we have bonds of love and friendship. I do not know the answer to this question, but I do think that it is important for normative ethicists to ask and answer it.

Other questions we must ask concern the future. For we can also imagine a whole series of ways the world could be in between the here-and-now and a time in the not-too-distant future, each only slightly dissimilar to the times adjacent to it, where the world at that later time is a brutal apocalypse. Moreover, there is a not insignificant chance that that imagined future is our future. The question we must ask ourselves now is whether there is anything we can do to ensure that that isn't our future, and if the not insignificant chance becomes significant with the passage of time, then the question we will have to ask ourselves then is how we can prepare ourselves for that future. Again, I do not know the answers to these questions, and I of course hope that we never have to ask that last question. But it would be both morally and non-morally wrong not to acknowledge the fact that that question may well be out there in our future waiting to be asked. It is a great virtue of the ethics of agency that it puts these questions about the brutal apocalypse on the normative ethics agenda.

References

Ayer, A.J. 1936: *Language, Truth and Logic* (London: Gollancz).

Calhoun, Cheshire. 2000. 'The Virtue of Civility' in *Philosophy and Public Affairs* (29) pp.251–75.

Falk, W. D. 'Guiding and Goading' in *Mind* (62) pp.145-171.

Hare, R.M. 1952: *The Language of Morals* (Oxford: Oxford University Press).

_____ 1972: 'The Argument from Received Opinion' in his *Essays on Philosophical Method* (Berkeley and Los Angeles: University of California Press).

_____ 1981: *Moral Thinking* (Oxford: Oxford University Press).

Jackson, Frank and Michael Smith 2006: "Absolutist Moral Theories and Uncertainty" in *Journal of Philosophy* (103) pp.267-283

Kelly, Thomas and Sarah McGrath 2010: 'Is Reflective Equilibrium Enough?' *Philosophical Perspectives* (24) pp. 325-359.

Kiesewetter, Benjamin. 2022. 'Are All Practical Reasons Based on Value?' *Oxford Studies in Metaethics, Volume 17* (Oxford: Oxford University Press) pp. 27–53.

Lipscomb, Benjamin 2021: *The Women Are Up to Something: How Elizabeth Anscombe, Philippa Foot, Mary Midgley, and Iris Murdoch Revolutionized Ethics* (New York: Oxford University Press).

Mill, John Stuart 1863: *Utilitarianism* (London: Parker, Son, and Bourn).

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

Paton, H. J. (1948). *The categorical imperative: A study in Kant's moral philosophy* (Chicago: University of Chicago Press).

Portmore, Douglas W. 2011: *Commonsense Consequentialism: Wherein Morality Meets Rationality*. (Oxford: Oxford University Press).

Rawls, John 1951: 'Outline of a Decision Procedure for Ethics', *Philosophical Review* (60) pp.177-97.

_____ 1974-5: 'The Independence of Moral Theory' in *Proceedings and Addresses of the American Philosophical Association* (48) pp. 5-22.

Singer, Peter 1979: *Practical Ethics* (Cambridge: Cambridge University Press).

Smith, Michael 2009: 'Two Kinds of Consequentialism' in *Philosophical Issues* (19) pp. 257-272.

_____ 2018: 'Constitutivism' in *Routledge Handbook of Metaethics* edited by Tristram McPherson and David Plunkett (London: Routledge) pp.371-384.

_____ 2020: 'The Modal Conception of Ideal Rational Agents: Objectively Ideal Not Merely Subjectively Ideal, Advisors not Exemplars, Agentially Concerned 16 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, 2020) pp.59-79

_____ 2021: 'The State of Nature' in *Oxford Studies in Normative Ethics* edited by Mark Timmons (Oxford: Oxford University Press) pp.270-289.

Strawson, P. F. 1962: 'Freedom and Resentment' in *Proceedings of the British Academy* (48) pp.187-211.

Williams, Bernard 1981: 'Internal and External Reasons' in his *Moral Luck* (Cambridge: Cambridge University Press) pp.101-13.

Wolf, Susan 1982: 'Moral Saints' in *Journal of Philosophy* (79) pp. 419-439