What's Fit for the Fallible

Richard Yetter Chappell*
University of Pennsylvania / Bowling Green State University

March 9, 2013

Abstract

What would a utilitarian agent look like? Some have taken the answer to describe an agent so incompetent and perverse that it casts doubt on utilitarianism itself. In this paper, I develop the strongest form of this "self-effacingness" objection to utilitarianism, based on the idea of a constitutive link between rationality and *normally* competent agency. Assuming this understanding of rationality for sake of argument, I then suggest two ways to defend utilitarianism. One appeals to a Railtonian 'sophisticated' or indirect utilitarian psychology, though I suggest some potential problems for this approach. The second involves showing how we can develop a direct utilitarian psychology within rational constraints. In the course of distinguishing these two alternative paths, I make a distinction between dispositions that are 'extrinsically desirable' and those that are desirable in virtue of being 'well calibrated for action' — a distinction that I then employ to illuminate the Gauthier-Parfit debate about whether it's rational to act on rationally desirable dispositions.

*Thanks to Errol Lord, Sarah McGrath, Tim Mulgan, Philip Pettit, Derek Shiller, Peter Singer, Michael Smith, and Helen Yetter-Chappell, for helpful comments and discussion.

Introduction

Bernard Williams (1973) famously objected that "utilitarianism's fate is to usher itself from the scene." The idea is that utilitarianism will tell us that we should try to rid ourselves of our belief in it (and perhaps internalize some alternative theory instead), since this would better achieve the goals of the theory.

Taking Williams' worry as a starting point, this paper will explore the relation between moral theories and morally fitting psychologies, with a particular focus on what's fitting for fallible, non-ideal agents. My main focus will be on utilitarianism, and whether the fitting utilitarian agent would be unacceptably incompetent in a way that casts doubt on the truth of the theory. But much of what I say in utilitarianism's defence could also be extended to Kantianism and other targets of Williams-style anti-theory objections.

I begin, in §1.1, by explicating the self-effacingness objection in greater detail, showing how it can be understood as challenging the utilitarian conception of a morally fitting agent, and why this 'fittingness objection' is a challenge to utilitarianism itself. §1.2 explains why the standard consequentialist response to character-based objections — namely, distinguishing criteria of rightness from decision procedures — is inadequate to meet the objection.

§2 explores the use of Railton (1984)'s 'sophisticated' consequentialist psychology in response to character-based objections. Insofar as the sophis-

ticated consequentialist agent possesses whatever motivations would be most desirable, it may seem that what proponents of this account are describing is merely a *fortunate*, not necessarily any kind of *fitting*, psychology. We can imagine circumstances in which it would be most fortunate to possess positively malicious motivations, but that clearly wouldn't make malice a virtue (or 'fitting', as I use the term here). So, I will argue, this popular approach also fails to adequately address the fittingness objection.

In diagnosing where the Railtonian conception goes wrong, I suggest that we can identify a proper *subset* of desirable dispositions as 'well-calibrated' or *rationality-enhancing*, in contrast to others that merely cause better results by some external means. It is only the rationality-enhancing dispositions that we should attribute to the fitting moral agent: Extrinsically desirable dispositions are worth pursuing, but once achieved may disqualify one's resulting psychology from being morally fitting. (Just as, for example, a hedonist should try to acquire non-instrumental interests of other kinds, but once they do they no longer qualify as fitting the theory of hedonism. They have rationally made themselves irrational, by hedonistic lights.)

As a brief aside, I take this distinction to illuminate the Gauthier-Parfit debate about rational transmission: i.e., whether it's always rational to act on a disposition that it's rational to acquire. In section §3, I argue that Parfit is correct that no such general principle holds: Even if the deterrence value renders it rational to acquire a doomsday disposition, it is not necessarily rational to *act* on such a disposition. Nonetheless, we may hope to vindicate

a related version of the transmission principle that is restricted to those dispositions that are independently identifiable as rationality-enhancing or 'well-calibrated'.

Finally, in §4, I draw on this conception of 'well-calibrated' dispositions to show how I think the utilitarian can successfully respond to the self-effacingness objection. In doing so, I will draw heavily on assumptions presupposed by the objection: namely, that there are norms fit for human-sized (finite, fallible) minds, and that a person who meets these norms is thereby competent to act in a wide range of circumstances. My general strategy is thus to first specify some initial preconditions for competent human-like agency, and then to explicate how a recognizably utilitarian mindset might fit within those constraints. The resulting picture, if still not entirely attractive, may at least seem significantly less "defective" — and less likely to be typically self-effacing — than the standard caricature of a utilitarian agent.

To this end, I will first show that the utilitarian agent should not be conceived of as constantly engaging in deliberate calculation. This misconception may arise from assuming that every deliberative question to which there is an appropriate answer is thereby a question that it's appropriate *to ask*, but we will see that this cannot be so for finite agents. Attention is a limited resource, and executive control should only intervene when by doing so it is likely to improve the quality of the agent's actions. *Excessive* executive control is, I will argue, no part of the fitting utilitarian mindset. So, the fitting utilitarian agent will not engage in predictably unreliable at-

tempts at explicitly calculating utilities, but will rely more heavily (as many utilitarian philosophers have suggested) on the general rules of thumb that are more likely to see him right. In particular, I will argue that the fitting utilitarian will *not* be disposed to break a generally beneficial rule merely because the benefits seem to him to outweigh the costs. Even on straightforward act utilitarian grounds, his behaviour may be largely rule-governed in a way that renders him trustworthy and eligible for social co-operation. Here I do not *merely* argue, as others have done, that this is the most fortunate or utilitarian-recommended mindset (indeed, depending on empirical contingencies, it may not be). Rather, I will argue that such a mindset is *fitting* to the utilitarian's theory, in light of rational norms for fallible agents.

1 Self-Effacingness

1.1 The Objection

The self-effacingness objection, read straightforwardly, objects to the mere fact that utilitarianism is *in fact* self-effacing. But this cannot reasonably be thought objectionable. To see why, first observe that any plausible moral theory is at least *possibly* self-effacing in this way, because any plausible moral theory will tell us to acquire false moral beliefs if this is the only way to avoid absolute disaster — and it is always possible to contrive situations in which this is the very choice that we face. So it cannot be objectionable that utilitarianism is merely possibly self-effacing. Moreover, since the true

moral theory is presumably non-contingent, it would be very odd to think that this objection suddenly gains additional force if we happen to live in one of those possible worlds where the theory is self-effacing. Moral facts cannot plausibly depend in this way on our location in modal space. So the mere fact that a moral theory turns out to be *actually* self-effacing is not objectionable either.

Next we may note that utilitarianism is not *necessarily* self-effacing: there are possible worlds where it is (expectably) best to believe it. So one cannot form a sound objection from that premise. But perhaps it would be genuinely objectionable if utilitarianism were, in some sense, *normally* self-effacing.[1] And if, as critics charge, the utilitarian agent is constantly calculating, untrustworthy, apt to break generally beneficial rules whenever it strikes him as optimal to do so, etc., then we may indeed expect the view to be quite typically detrimental when internalized by fallible, human-like agents. The utilitarian mindset begins to look not just *unfortunate* ("given the circumstances"), but intrinsically defective.

This, then, is what I take to be the strongest interpretation of the self-effacingness objection. First, the critic assumes a strong connection between rationality and (metaphysically, not statistically) *normal*[2] competence, such

[1] Williams himself refers to "empirical generalities of a kind which are the background to all problems of morality" (1973, 134), but it isn't clear that he really means his argument to be interpreted in the way I recommend below. Nonetheless, I propose that this is the most interesting objection in the vicinity, whether or not it is what Williams intended.

[2] This is the sense in which it is normal for cats to have four legs, even if every actually existing cat is an amputee (so that the median number of cat legs is three or fewer). This clarification secures the modal robustness of rational norms: It shouldn't turn out that

that any normally incompetent agent is ipso facto lacking adequate rational capacities. Second, we need a link between moral agency and rational agency, such that virtue, or moral fittingness, is not incompatible with the possession of adequate rational capacities. Third, the critic posits that any fitting utilitarian agent would be normally incompetent. From this he draws his first conclusion: that the fitting utilitarian agent is not morally fitting. Add the following conceptual truth: if an agent that 'fits' some theory X is not yet *morally* fitting, then X is not the true moral theory. Now we can derive the critic's conclusion: Utilitarianism is a false moral theory.

In this paper, I grant the first two premises for sake of argument. I seek to defend utilitarianism by instead rebutting the third premise: that a fitting utilitarian agent would be normally incompetent. Critics have thought this because they imagine the utilitarian agent as one who explicitly makes an expected utility calculation before each decision; who finds the needs of those before his eyes to be no more salient than those inaccessible and far away; and who is ready and willing to commit atrocities in the name of efficiency, without hesitation or regret. But, I will argue, these assumptions are mistaken. This is not what a fitting utilitarian agent would look like, as

the fundamental norms of rationality differ from world to world. But which psychological dispositions are statistically most *often* useful is something that radically differs from world to world, depending on external contingencies. So if we are to tie rationality to normal competence, we cannot mean 'normal' in the statistical sense. Rather, 'normality' in the relevant sense is something that is held fixed regardless of actual-world contingencies. One (controversial) way to precisify the notion is to assume that there is an *a priori* objective *probability distribution* over the possible worlds, and the 'normal' worlds are those that had the highest a priori probability of being actualized.

we can see when we take into account rational norms for fallible agents with human-sized minds.

Before we move on, I should say a little more to clarify the sense of 'fittingness' that features in this argument. This is a term of art that is meant to capture the intuitive idea of what's warranted or rational *from the perspective of* a moral theory — or, when used in a non-relational context, what's warranted or rational from the point of view of the *true* moral theory. It's fitting to desire that which is good or *desirable*, to admire the admirable, and so on. So, for example, if utilitarianism holds that what's good is just the welfare of sentient beings, then the fitting utilitarian agent is one who desires just the welfare of sentient beings. If such desires are shown to be not actually fitting, then that's just to say that utilitarianism is false: it makes mistaken claims about which things are desirable.

Of course, what's rational or fitting to desire may come apart from what it would be *best* or most fortunate to desire, just as what it's rational to believe (based on the evidence) may come apart from what it would be best or most fortunate to believe (given various practical incentives). So utilitarians can comfortably say that we 'ought', in the practical sense, to have whatever desires would be most fortunate, without thereby committing themselves to the view that those desires are *fitting*, or that their objects are truly *good*. And indeed, utilitarians have traditionally been much more interested in the question of which desires are best to have, than that of which are rational or fitting (according to their theory). But neglect of the latter question does

not mean that there isn't a real question there to be asked. And, as we'll see, it's a question that the utilitarian *needs* to answer if they are to offer an adequate response to the argument laid out above.

1.2 Why the standard response fails

Consequentialists standardly distinguish between *criteria of rightness* and *decision procedures* (Bales 1971). Just because utilitarians hold that an act is right iff it maximizes expected utility (say), it doesn't follow that they recommend actually trying to calculate utilities in your everyday life. Indeed, given that such constant calculation would be predictably counterproductive (due to lack of time, misleading evidence, cognitive bias, setting bad precedents, etc. — see Mackie 1985), utilitarians would strongly recommend against it!

All this is true enough, but besides the point. The objection is not to utilitarianism's *recommendations*, but to its *implications*. There's a fact of the matter as to what the 'fitting' utilitarian psychology is, quite independently of what psychology utilitarianism *recommendeds* that we try to inculcate. But if the fitting-utilitarian psychology can be shown to be not actually morally fitting, that would — as previously explained — entail the falsity of utilitarianism as a moral theory.

Any moral theory has some implications for what kind of psychology is 'fitting' or rational from a moral point of view. At a minimum, as we've seen, one's theory of the good commits one to holding certain objects (namely, the

things identified as good) to be *fitting to desire*. Again, that's not to say that a utilitarian must recommend that we try to acquire fitting desires — it's an open empirical question whether being 'rational' in this way would promote utility in actual circumstances. So that's not the issue. Rather, the issue at hand is just whether the kind of mindset utilitarianism *implies* is rationally fitting *really is* so. If it's not, then the theory is shown to be false, in virtue of having false implications. For example, if utilitarianism implied that people are 'replaceable', in the sense that it's rationally fitting to desire each person's welfare merely as a means to promoting the aggregate welfare, then (assuming we're right to doubt the latter claim) that would be grounds for thinking that the theory must be mistaken.[3]

So we can't just ignore decision-procedures and other psychological elements. And nor can we merely settle for identifying those which are most conducive to utility, and thus *recommended* by utilitarianism. As normative theorists, interested in whether or not utilitarianism is a true moral theory, we must also investigate what kind of mindset would be a *rationally fitting* mindset, were utilitarianism true. We can then test whether this fitting utilitarian mindset meets the minimal requirements for rationality, such as the 'normal competence' test proposed in §1.1, and hence whether utilitarianism itself remains an eligible moral theory.

In the following sections, I explore two very different strategies for con-

[3] Happily, as argued in Chappell (forthcoming), utilitarianism does *not* imply such replaceability!

structing a non-defective utilitarian psychology in answer to this challenge. §2 explores the Railtonian "sophisticated" psychology, with non-utilitarian desires. §4 sets out my preferred "subjective" account, arguing that critics are mistaken to assume that a fitting agent with utilitarian motivations would be guided by explicit "expected utility" calculations.

2 Sophisticated Utilitarianism

2.1 Explication

Railton (1984) contrasts two kinds of hedonistic (or, more broadly, consequentialist) psychologies: 'subjective' and 'sophisticated'.[4] The subjective hedonist is solely motivated by concern for his own happiness. However, the 'paradox of hedonism' suggests that such a person is likely to end up quite unhappy. Happiness may be better achieved by those who are motivated by other concerns. Railton thus introduces the sophisticated hedonist — let's call her 'Sophie' — who "aims to lead an objectively hedonistic life (that is, the happiest life available to [her] in the circumstances) and yet is not committed to subjective hedonism." Sophie may thus possess and act on distinctively non-hedonistic motives — e.g., concern for others — if such desires are conducive to her living a happier life overall.

Once she has moved beyond subjective hedonism, and acquired a happy collection of non-hedonistic motivations, we may begin to wonder in what

[4] Below I follow Railton in focusing on hedonism for simplicity. But everything I say should carry over, in obvious fashion, to the case of utilitarianism.

sense Sophie is still a "hedonist" at all, rather than a whole-hearted pluralist. What sets Sophie apart, according to Railton, is that her psychology continues to be regulated by a **counterfactual condition** according to which, despite her various desires, she "would not act as [s]he does if it were not compatible with [her] leading an objectively hedonistic life."

This requires some unpacking. One might be tempted to interpret the counterfactual condition as applying at the level of individual acts, so that Sophie always chooses the hedonistically recommended action. But this would seem to require an overriding desire for happiness. The most that can be said for Sophie, on this interpretation, is that her non-hedonistic desire may be the one that is causally operative in cases of motivational overdetermination.[5] But her hedonistic desire is always there, waiting in the wings, ready to swoop in and exert its overriding force whenever it appears that she is about to make something other than the hedonistically-recommended choice. Sophie thus looks little different from the subjective hedonist, on this interpretation.

Alternatively, we may interpret the counterfactual condition as applying at a more global level (as suggested by Railton's reference to the "objectively hedonistic *life*"). Whereas the subjective hedonist regulates her individual actions according to hedonistic norms, Sophie's hedonism instead regulates her desires and dispositions. So, for example, her pro-friendship disposition may lead Sophie to perform individual acts that reduce her happiness — e.g.

[5] I owe this suggested interpretation to Michael Smith.

answering her friend's distraught 3 a.m. call — but her 'hedonic monitor' is not triggered to intervene unless it becomes clear that the relationship *as a whole* is detrimental to her happiness, such that she would be better off in the long run with different desires and dispositions.

Some questions remain concerning the precise details of how this regulative mechanism is supposed to work. In particular, we may wonder whether Sophie has an overriding desire to *possess hedonically fortunate dispositions*, that she will act upon (overriding her other, non-hedonistic motivations) whenever she's in a position to do so. This may still seem too close in structure to the psychology of the simple consequentialist. It so happens that Sophie's overriding desire, due to its quirky content, applies to fewer situations than does the corresponding desire of the subjective hedonist. (More of our choices potentially impact our happiness than potentially impact whether we have happiness-promoting dispositions.) But this looks more like a difference in degree than a difference in kind. Indeed, in light of the structural parity, such an agent may be better described as a straightforward subjective maximizer of happiness-promoting dispositions, rather than a sophisticated maximizer of happiness!

We may do better to interpret Sophie's hedonism as manifested not in a *desire* at all, but rather a higher-order mechanism that serves to regulate her desires through some sub-personal causal process. The key difference is that this hedonistic mechanism, unlike a desire, never directly manifests itself in action. It is not *itself* a motivation that she may act on (though it may cause

her to acquire some independently motivating hedonistic desires, insofar as these would cause her to live a happier life). Its control over her actions is instead wholly indirect: The hedonic monitor shapes Sophie's desires in hedonically fortunate ways, and then she acts on *these desires*, whatever they may be.

One worry for this view is that the hedonistic desire-regulating faculty begins to look like an 'external' imposition on Sophie — turning her into a kind of puppet. To avoid this problem, we need to ensure that the regulating faculty is in some sense under Sophie's rational control. It is *because* Sophie believes in (or has a deep-rooted standing commitment to) hedonism that the faculty regulates her desires according to hedonistic goals. If she came to fully believe and endorse some other normative theory, say utilitarianism, then the faculty would regulate her desires according to utilitarian goals instead. The 'sophisticated' psychology is thus best described in two parts: First, there is the agent's overarching "primary goal", which she may identify with during reflective moments, but which does not tend to directly motivate her actions. Instead, she is moved by the "secondary" desires and dispositions that are produced and regulated by a mechanism that is responsive to her primary goal.

2.2 Evaluation

Supposing that the psychology described in §2.1 is coherent, it's an interesting question how exactly we should evaluate it. Suppose that (egostic)

hedonism is true, so that one's own pleasure is the only end that's truly desirable, or worth pursuing. Sophie then seems irrational, in that her desires do not conform to the normative facts about what's desirable, and her actions likewise fail to be sensitive to hedonistic reasons: She often benefits others at her own expense. On the other hand, she is not *completely* insensitive to reasons: Her desire-regulating faculty ensures that she maintains the desires that (the evidence suggests) it is best for her to have — and if circumstances change, so will her dispositions. This suggests an important sense in which Sophie's reflective hedonism is ultimately 'in control', even if it is not what moves her. We may thus need to draw a distinction between (local) act and (global) agent rationality, allowing us to say that *Sophie* is rationally fitting or responsive to reasons, even if her particular *actions* are not.

It's worth noting that even this vestige of rational sensitivity may, in special circumstances, make her worse off. Consider Parfit (1984)'s example of the society of perfectly rational egoists, some of whom come to realize that it will advance their interests to become irrational in a specific respect: namely, if they become transparently disposed to follow through on their threats regardless of the costs to themselves. Such a "threat-fulfiller" can then strap a bomb to his chest, and threaten an egoist that he will detonate it (killing them both) unless the egoist complies with his whims. He can safely make such threats, because he knows the egoist would sooner comply than die. As Parfit further shows, the rational response for the remaining egoists is to turn themselves into transparent "threat ignorers", who are stably dis-

posed to (irrationally) ignore threats no matter the costs to themselves. A threat-fulfiller will leave the ignorers alone, because he knows that if he were to threaten them, they would ignore him, and he would then detonate the bomb, killing them both. (Note that the threat-fulfiller will not *issue* threats that he expects will make him worse off. It is merely *fulfilling* threats that he does blindly.)

In comparison to the pure threat-ignorers, Sophie is more apt to have her rationality exploited. Given transparency, the threat-fulfiller will know that if he threatens Sophie, she will comply. For Sophie's regulating mechanisms will not allow her to maintain a disposition once it becomes clear that it is disastrous for her long-term happiness. And a threat-ignoring disposition becomes clearly disastrous as soon one is actually issued with a credible apocalyptic threat. So, a threat-fulfiller will know that he can safely threaten Sophie, and she will (if necessary change her dispositions and) comply rather than die. To avoid such exploitation, Sophie would have to alter her psychology so that she would become a *pure* (unregulated, insensitive) threat-ignorer — at which point she would no longer be a sophisticated hedonist. She would just be (however fortunately) plain irrational.

We thus find that a Railtonian sophisticated hedonist (or, more broadly, sophisticated consequentialist) psychology is by no means guaranteed to endorse itself as the most fortunate psychology to possess in every possible situation. But it offers a suggestive alternative to the standard conception of a rational psychology. Insofar as we are drawn to the idea that rationality

should not *normally* be a curse (even if it may be in certain special circumstances), we may see Sophie's two-level psychology — with its capacity for her primary goal to control and regulate her secondary, action-guiding motivations — as an improvement over the subjective hedonist's unitary motivational structure. While acknowledging that Sophie's actions are often locally irrational (by hedonist lights), we may be more concerned to evaluate her global rationality as an agent. In this respect, at least, she may at first glance seem more reasonable.

I think there are important grounds for doubting this conclusion, however. Let's return our attention from hedonism to utilitarianism. The sophisticated utilitarian — call her 'Sophu' — will have whatever motivations are most conducive to promoting the general welfare. So, in particular, if an evil demon threatens to torture an innocent population unless Sophu comes to intrinsically *want*[6] them to suffer, then Sophu will be led to acquire this fortunate but malicious motivation.[7] This is a good outcome, in the cir-

[6] One might wonder if Sophu's resulting pro-suffering desire will be merely instrumental, since it is produced (by her regulating mechanism) as a means to promoting welfare. But we must take care to distinguish conditions on the desire's *existence* from conditions on its *content* (or, in other words, to distinguish a desire's persistence conditions from its fulfilment conditions), and I take the instrumental/intrinsic distinction to concern the latter. It may be that Sophu's regulating mechanism ensures that the desire's existence is *contingent* upon its promoting utility, but that doesn't mean that considerations of utility enter into the *content* of the desire. What Sophu ends up wanting — i.e., the content or object of her desire — is simply *that people suffer*, not some other end to which this suffering is a means. She becomes motivated to pursue suffering for its own sake. It's just that her motivations will change when they cease to promote (expected) utility.

[7] At least, this is so on my interpretation of the 'sophisticated' psychology. Mason (1999, 256) suggests an alternative view on which "We can develop new motives from old motives, but only when they are consistent." This would rule out a sophisticated utilitarian acquiring malicious motivations, however beneficial they might be. However,

cumstances, as it prevents a lot of suffering. But if any desire is held to be irrational or unfitting by utilitarian lights, it is surely an intrinsic desire that others suffer. Sophu has, quite virtuously, made herself vicious by utilitarian lights. And note that it is not just her *actions*, but her *desires* — her very *self*, we might think — that is impugned here.

The advocate of sophisticated utilitarianism might at this point defend Sophu by pointing out that her *deepest commitments* remain pure and altruistic, even as they respond to the unfortunate circumstances by shaping her motivations in this malicious-but-instrumentally-valuable direction. So there at least remains *something* fitting about Sophu's psychology. The tricky question, which I will not resolve here, is whether this is enough to allow her to qualify as a fitting utilitarian agent *overall*.

3 Rational Transmission

It seems plausible to think that there's a tight connection between (i) the rationality of acquiring and maintaining a disposition, and (ii) the rationality of 'acting on' the disposition, i.e. performing an action that the disposition characteristically disposes you towards. One candidate connection is suggested in the following simple principle of rational transmission:

(RT-past) For any disposition D and act A that is characteristic

it's not clear to me whether this is meant to be a conceptual constraint on rational agency, or just a contingent empirical hypothesis about how new motivations actually develop in people.

> of D: *If it was rational to acquire D then it is rational to perform A.*

But Parfit (1984)'s above-described case of the threat-fulfillers casts doubt on this principle. It may well be rational for a self-interested agent to acquire the threat-fulfilling disposition, but if (through some irrational quirk) a threatened target unexpectedly ignores the agent's apocalyptic threat, it is surely *not* rational for the agent to follow through and blow themselves up. Such disastrous stubbornness would seem, on the contrary, quite crazy.

Gauthier (1997) is not wholly convinced by this counterexample to RT-past, but suggests and endorses two weaker transmission principles, which we may formulate as follows for any disposition D and act A that is characteristic of D:

> **(RT-timeless)** If it was rational to acquire D *and is better to maintain D than never to have possessed it at all*, then it is rational to perform A.
>
> **(RT-present)** If it was rational to acquire D *and is rational to maintain it presently*, then it is rational to perform A.

However, Parfit (2011, Appendix B) points out that even these weakened transmission principles are susceptible to counterexamples, such as:

> *Schelling's Case.* A robber threatens that, unless I unlock my safe and give him all my money, he will start to kill my children.

> It would be irrational for me to ignore this robber's threat. But even if I gave in to his threat, there is a risk that he will kill us all, to reduce his chance of being caught. [... I]t would be rational for me to take a drug that would make me very irrational. The robber would then see that it was pointless to threaten me; and since he could not commit his crime, and I would not be capable of calling the police, he would also be less likely to kill either me or my children. [...] But while I am in my drug-induced state, and before the robber leaves, I act in damaging and self-defeating ways. I beat my children because I love them. I burn my manuscripts because I want to preserve them.

Parfit stipulates that these destructive acts are not necessary to convince the robber that you are irrational. So they have no good effects, though they stem from a disposition that it is worthwhile (for extrinsic reasons) to possess. Are these acts rational? I share Parfit's sense that they are not. So all of the transmission principles surveyed thus far fail.

The fundamental explanation for this disconnect is that an agent's dispositions can have other consequences besides producing downstream acts in the agent herself. In particular, you might be harmed or rewarded directly on the basis of whether you possess some disposition, independently of whether you act on it. This suggests that we can distinguish (i) dispositions that have high expected value, all things considered, and (ii) dispositions that have high expected value *in respect of the downstream actions they'll tend to*

produce. We can call the former class of dispositions 'desirable', and the latter 'well-calibrated'. Dispositions that are desirable but *not* well-calibrated we may call 'extrinsically desirable'. It is these extrinsically desirable dispositions that feature in Parfit's cases of 'rational irrationality', i.e. whereby it is rational to acquire and maintain such a disposition, but not to act upon it.[8]

While acknowledging this possibility, we may still think that there must be *some* 'transmission' principles according to which the rational status of a general rule or disposition can be inherited by the particular acts it prescribes. And, indeed, the distinction I've just highlighted suggests an obvious candidate principle: we just need to restrict the dispositions in question to those that are 'well-calibrated', i.e. desirable for their (expected) impact on your downstream actions, rather than for extrinsic reasons. Consider the following transmission principle:

> **(RT-Calibrated)** For any dispositional set D and act A that is characteristic of D: *If D is* well-calibrated*, i.e. expectably good to possess in virtue of the downstream actions it tends to produce, then it is rational to perform A.*

Alternatively, we may formulate the principle in terms of rules rather than dispositions:

[8] This parallels the familiar distinction between 'object-given' and 'state-given' reasons. The dispositional state is a useful state to be in, but the (metaphorical) 'object' of the state — the set of actions it disposes you towards — does not merit such a disposition.

> **(RT-CalibratedRule)** If S rightly adopts a rule R as maximally well-calibrated (i.e. expectably better to internalize than any conflicting rule, in virtue of the downstream actions it prescribes) and R prescribes ϕ-ing in circumstance C, then when S is in circumstance C, S rationally ought to ϕ.

Such a principle may be supported by considerations of meta-coherence. Ex hypothesi, the rule R offers the most reliable guidance available to S — in particular, it produces good actions more reliably than does attempting to autonomously determine what the best result would be in each case. (And it is also more reliable than any identifiable alternative, e.g. "following R except in circumstances with the subjectively distinguishing feature F.") So, any given departure from R can be expected to have worse results than would be obtained by following R. So in any given case, the agent should follow R's advice.

Actually, this isn't quite right. The argument from metacoherence only supports acting on rules that improve the agent's actions due to serving as an *epistemic guide*, like the rule against killing people even when murder might (prima facie) *seem* to promote utility. But in non-ideal agents, a disposition might also conduce to good actions for a very different kind of reason, as I will illustrate below.

Meet Cam, a callous consequentialist. Cam is one of those utilitarians who likes humanity but not people so much.[9] Due to his lack of regard for those around him, he tends to act insensitively, and makes other people (not least his poor family) miserable. Upon reflection, Cam recognizes this to be unfortunate. But he lacks the strength of will to reliably act better in such situations. So he takes a pill that makes him a much more caring and loving person. He is now disposed to attend disproportionately to the welfare of those that are close to him. This causes him to act in much better ways: in particular, he finds it easier to refrain from making the kinds of insensitive remarks that previously caused so much harm. The one downside is that he is now much less inclined than before to donate to GiveWell-recommended charities that promote the impartial good. He would rather spend that money on his family. This is harmful, but (let's suppose, perhaps unrealistically) not nearly as harmful as Cam's callous actions had been.

The structure of the case is that Cam was previously *weak-willed* in a very bad way. He then acquired a disposition that helped him to overcome this weak will, and so perform better actions — though at the cost of acquiring a new (less bad) weakness. Because it is clear (by stipulation) which of his newly-disposed actions are better than before and which are worse, the epistemic argument for following a generally beneficial disposition no longer applies. (The overall benefits of the change aren't evidence that his giving less

[9] Just to clarify: This nasty streak is not, of course, any part of the fitting utilitarian psychology!

to charity, in particular, is beneficial. On the contrary: *this* action remains clearly bad, it's just that value of his *other* actions outweighs this localized badness.) So, we find that it's right for Cam to acquire this disposition in virtue of the actions it conduces to in general, but some particular actions it conduces to may still be considered wrong, or less than perfectly rational from a utilitarian perspective. We can thus have cases of what Parfit (1984) would call 'blameless wrongdoing', without having to appeal to dispositions that have good effects besides action. (Of course, the disposition may also be good for other reasons — the point is just that my case doesn't rely on this.)

Perhaps we can get around such cases by restricting the transmission principle to dispositions that are *maximally* well-calibrated:

> **(RT-CalibratedMax)** For any dispositional set D and act A that is characteristic of D: *If D is* maximally *well-calibrated, i.e. the expectably best dispositional set to possess in virtue of the downstream actions it tends to produce, then it is rational to perform A.*

This assumes that there is a possible dispositional set that achieves the best of both worlds, relieving Cam of his previous character flaws without adding any new ones. But we may doubt whether that is always possible, and it also restricts the usefulness of the principle for non-ideal agents. Perhaps the best we can say is that when a rule or disposition is well-calibrated for action

in virtue of serving as an epistemic guide, then following the guidance of the rule or disposition is rational.

A deeper cause for hesitation comes from considering cases where you adopt a rule as a hedge against (as it happens, misleading) evidence that you might be biased in your subsequent judgments. For example, the rule might tell you to disregard certain first-order evidence, because you can't be trusted to evaluate it rationally. But if you actually *are* capable of evaluating it rationally, then we may think that there's an important sense in which you rationally ought to be guided by the (first order) evidence. Or, even if the higher-order evidence makes some contribution, we may still doubt the radical claim that it *completely swamps* the first-order evidence in determining what you rationally ought to do (cf. Kelly 2010). In that case we may similarly doubt that there are any true and interesting (non-trivial) principles of rational transmission from dispositions to individual acts.

In sum: While I am uncertain that any such principle is ultimately vindicated, focusing on the subset of dispositions that are *well-calibrated*, in my described sense, would seem to give us the best shot. And, as we will see, these are just the dispositions that may be possessed by the "subjective" act utilitarian agent, in contrast to the unfitting but (extrinsically) desirable dispositions that we saw could be part of the "sophisticated" utilitarian psychology.

4 The Act Utilitarian Agent

Suppose we accept my earlier suspicion that "sophisticated" psychologies, with their extrinsically desirable dispositions, are not rationally fitting psychologies. The remaining option for defending utilitarianism against the argument of §1.1, is to spell out a non-defective "subjective" utilitarian psychology. In this section, I will attempt this task, by drawing on the critic's assumption that there are rational norms for human-sized minds that render an agent competent to act in normal circumstances. I will especially make use of the idea that our account of the fitting utilitarian agent, while restricted to utilitarian motivations, may at least appeal to *well-calibrated*, if not merely extrinsically desirable, guiding dispositions. This restriction is one of the main features that sets apart my straightforward account of the fitting utilitarian psychology from the "sophisticated" view explored in §2.

4.1 Motivating vs. Guiding Dispositions

Let's begin by distinguishing what I'll call 'guiding' and 'motivating' dispositions.[10] Our non-instrumental desires or *motivations* are our driving concerns, or what move us to action. They represent the goals we hope to realize through acting. On the other hand, this motivational 'oomph' can be steered or *guided* by strategies and heuristic dispositions that shape our behavioural responses in pursuit of those goals. We may think of our guiding dispositions as, roughly, the psychological manifestation of instrumental

[10] Thanks to Philip Pettit for his assistance in formulating this distinction.

rationality. They take our desires as inputs, and output a suitable action or intention.[11]

The standard caricature of the utilitarian agent assumes that we can "read off" both kinds of dispositions from the moral theory. From its theory of the good — the view that what matters is just the welfare of sentient beings — we get the fitting utilitarian motivations. That much I agree with: the fitting utilitarian will desire the welfare of sentient beings.[12] But the standard caricature also takes the 'maximizing' aspect of utilitarianism to settle the *guiding* dispositions of the fitting utilitarian agent: they will (allegedly) decide how to act by, in each instance, conducting an expected-utility calculation, and then perform whatever action they judge to have the highest expected utility. It is this feature of the imagined utilitarian agent that is responsible for so much of their apparent defectiveness. And it is this feature that I deny we should attribute to the fitting utilitarian agent.

Instead, I propose that our choice of moral theory only commits us to the fittingness of the corresponding *motivating* dispositions. When it comes to the fittingness of guiding dispositions, this is instead determined by our independent — morally neutral — account of instrumental rationality. Drawing on our critic's understanding of *rationality as normal competence*, we can elucidate the fitting guiding dispositions as those that are prerequisites for

[11] I remain neutral on the question of whether practical reasoning is best understood as concluding in action or intention.

[12] Though it's an important question whether we interpret this as a single monolithic desire for aggregrate welfare, or — as I prefer — a plurality of desires, one for *each* sentient being's welfare. See my (forthcoming).

normally competent agency in creatures with human-sized minds.

My strategy for responding to the self-effacingness objection of §1.1 is thus as follows. I will offer a brief sketch of some 'well-calibrated' guiding dispositions which I take to be (a) prerequisites for normally competent human-like agency, and hence (b) rationally fitting for agents with human-sized minds. I will then show how an agent with fitting utilitarian motivations could also possess these well-calibrated guiding dispositions. The result is a non-defective, normally competent, fitting utilitarian agent. I will wrap up by illustrating how my 'well-calibrated' fitting utilitarian can be used to address several prominent character-based objections to utilitarianism.

4.2 Defective Deliberation and the Well-Calibrated Agent

Let me spell out four central features of the fitting agent's guiding dispositions. Firstly — as perhaps the most obvious prerequisite for competent agency — we have *epistemic rationality*: that is, the agent must have well-calibrated expectations about their environment. They cannot take the roar of a dangerous predator as evidence that a cute puppy awaits them outside. They need to have generally reasonable beliefs about their environment, and about what would be effective means for realizing their ends (whatever those might be).

Next, at the borderline of the epistemic and the practical, we will find constraints on how the agent is disposed to allocate their limited *attentional* resources. They must be generally attentive to possible threats and oppor-

tunities in their immediate environment, while also — in a calm moment, when appropriate — considering more abstract mental models of past and possible future scenarios (for sake of planning, self-evaluation, etc.). The details aren't too crucial for my purposes, but as we'll see, it's important that the fitting agent not dwell excessively on the past.

Thirdly, the competent agent requires well-calibrated habits, instincts, or sub-personal "predispositions" (Pettit and Brennan 1986) — an "auto-pilot" set, e.g., to avoid pain, be cooperative, and help others in need — to secure effective automatic behaviour in normal circumstances. One reason for this is that in time-critical situations, the agent cannot afford to pause to reflect on their situation at all. Often, a competent agent will be moved immediately (without conscious deliberation) to act, upon registering pertinent information about their environment. This is no mere behavioural reflex, as the agent is genuinely acting for reasons. But the rational processing goes on 'below the surface', as it were.

Once equipped with such well-calibrated predispositions, the fitting agent may act on them without need for excessive self-monitoring or executive control, and — in so doing — they may trust that they are acting for the best. Our fitting agent may, in this way, reap the practical benefits of 'satisficing' without the theoretical baggage.[13]

The fourth and final component that I'll discuss here is the possession

[13] Cf. Slote and Pettit (1984). Slote's satisficing consequentialist merely aims to achieve 'good enough' consequences, which makes sense as a practical strategy but seems rather more puzzling as an account of the rationally warranted *ultimate aims*.

of well-calibrated *triggers* for executive oversight. On pain of regress, we cannot deliberate about whether to deliberate. So, as previously noted, the agent's *default* guidance must be from non-deliberative "predispositions". But when these are not up to the task — when, say, the agent is faced with novel or complex circumstances for which their predispositions aren't so well calibrated to deal with — the agent's sub-personal mechanisms must recognize this and respond by triggering explicit deliberation on the part of the agent.

In summary: the fitting human-like agent — if they are to be capable of acting competently in a wide range of 'normal' circumstances — will rely heavily on well-calibrated predispositions, rather than explicit deliberation or calculation, to guide their actions in pursuit of whatever their goals may be. And this will be so even if their goal is to promote the well-being of sentient creatures as much as they are able. This, I propose, is how we should understand the fitting utilitarian agent. They have straightforwardly utilitarian desires, which are then translated into action via the above-described 'well-calibrated' guiding dispositions.

4.3 Addressing The Objections

We are now in a position to assess how my conception of the fitting utilitarian agent stands up to various anti-utilitarian objections.

We can first note that my 'well-calibrated' fitting utilitarian will not be "constantly calculating". Absent any triggering of their executive faculty,

the fitting utilitarian will respond directly to the salient needs of others — a child drowning in a pond, say — without mediation by explicit deliberation, let alone abstract judgments of "permissibility". In this way, the fitting utilitarian will not exhibit what Williams (1982) famously called "one thought too many".[14]

The fitting utilitarian's reliance on generally-reliable predispositions also undermines the objection that they would engage in "marginally-beneficial rule-breaking", such as breaking a promise whenever the benefits from doing so seem to even slightly outweigh the costs.[15]

Because overt calculation often goes awry, the competent utilitarian will — as we've seen — rely heavily on her generally reliable predispositions in everyday life, only pausing to reflect when her well-calibrated sub-personal mechanisms alert her to the need (say due to complex novel circumstances, that her "auto-pilot" wasn't designed to deal with). Everyday promise-keeping is not exactly novel, so for the fitting agent the question whether to keep a promise *shouldn't even arise*, unless there's something special about the situation that calls for her executive oversight.

That's enough to defeat the claim that the fitting utilitarian would com-

[14] I should mention that the traditional worry here is not just that the extra thought will make the agent too slow to act, but that it reveals something distressingly 'alienated' about his psychology, and the seemingly 'instrumental' nature of his concern for others. I further address such concerns in my second chapter, 'The Fitting and the Fortunate'.)

[15] Hooker (2000) raises the objection in terms of act utilitarianism implying that it's *objectively right* to break the rules if the benefits of doing so actually marginally outweigh the costs. I don't have particularly strong intuitions about objective rightness, as opposed to *what a moral agent would do*, so I restate the objection here in the latter terms.

monly engage in marginally-beneficial rule-breaking. But we may draw an even stronger conclusion. For suppose that our agent's executive oversight happens to be triggered. In a typical case, what should she conclude? We can stipulate that in fact the outcome would be marginally better if she broke her promise, but presumably the agent herself will not have any easy way of knowing this. (Among other things, she'd need to first consider the possibility of self-serving bias corrupting her judgment, and also to weigh the apparent benefits of rule-breaking in this instance against the long-run value of retaining a reputation for trustworthiness.) Maybe if she heard the booming voice of God reassuring her of this fact, then she could rationally go ahead and break her promise without further worry. But in *ordinary* circumstances — as we're supposed to be concerned with here — it's almost never going to be *clear* that rule-breaking is beneficial unless it is significantly (not merely marginally) so.[16]

So our agent is faced with an immediate choice: she can (i) break the rule even though it's not yet clear to her whether this would have good results on net; (ii) sink further cognitive resources into investigating a question that she probably shouldn't have bothered to ask in the first place; or (iii) simply keep her promise and turn her attention to more important matters. It seems pretty clear that, in this sort of case, option (iii) is the way to go.[17]

[16] Moore (1903) claimed that agents will never be in an epistemic position to justfiably break such rules, but we needn't be quite so pessimistic.

[17] In special circumstances, option (ii) may be truly costless, and so there's a possibility that the agent could reasonably undertake such an investigation, and responsibly reach the true conclusion that breaking the rule really is justified in this case. But this won't

In sum: Breaking a rule will generally only be obviously worthwhile in cases where it is also of significant benefit (in which case many would approve of rule-breaking anyway). If it's only of marginal benefit, this fact typically won't be sufficiently clear for a reasonably self-doubting, fallible agent to immediately act upon. And the low potential payoff means that it isn't really worth inquiring further: better just to stick with the generally-reliable rule of thumb. So a rational utilitarian generally won't be found engaging in marginally beneficial rule-breaking after all. (They'd even share our intuition that there's something awfully dubious about any agent who would act that way.) This gives them the kind of stable predictability needed for others to regard them as eligible and (more or less) trustworthy partners for social cooperation.

This discussion brings out the fact that the standard caricature of a utilitarian agent assumes that they will be unreasonably overconfident in their ability to calculate utilities accurately. But even if a utilitarian *initially* judges (just based on the first order evidence) that they would do best to break some generally beneficial rule, they may also realize that most people who make such judgments in similar situations are mistaken. Since they have no particular reason to think that they are one of the lucky few who make this judgment correctly, the general fact serves as a kind of higher-order evidence that their initial judgment was mistaken. All things considered, then,

be typical, and the crucial point for my purposes is just that one's *prima facie* utility judgments won't provide sufficient justification for breaking generally-reliable rules.

a reasonable expected utility judgment should, in this sort of circumstance, end up reinforcing the general rule rather than licensing typically-misguided unilateral rule-breaking.

The objections considered thus far — that the utilitarian would have "one thought too many", and that they would engage in "marginally-beneficial rule-breaking" — suggest the need to distinguish (i) the appropriate answer to a question, and (ii) whether a well-functioning agent would ask that question in the first place. The need for this distinction becomes especially apparent when we consider the following objection from Stocker (1989, 321):

> Maximizers hold that the absence of any attainable good is, as such, bad, and that a life that lacks such a good is therefore lacking. I disagree. One central reason for my disagreement stems from the moral psychological import of regretting the absence or lack of any and every attainable good. This regret is a central characterizing feature of narcissistic, grandiose, and other defective selves. It is also characteristic of those who are too hard on themselves, who are too driven and too perfectionistic.

This objection seems to me misguided. It may be unfortunate, and indeed even inappropriate ("defective"), to actively regret every little regrettable thing. But those things may be regrettable all the same. Crucially, this is not to say that a rational agent must regret them. It is more like a hypothetical imperative: *if* you closely attend to the features in question, this should induce in you feelings of regret. But it may be a kind of rational defect to

attend to the wrong things, if there are more pressing matters to attend to. As we saw in §4.2, the fitting agent would allocate their attentional resources in a way that avoids excessive dwelling on hypotheticals. So we can agree with Stocker that the agents he describes are defective, without thinking that the maximizing utilitarian would exhibit any such trait. On my picture, the utilitarian will have only a conditional disposition to regret the lack of a good *insofar as she attends to this lack*. But she'll usually have more important things to attend to, so she shouldn't actually end up actively regretting things very often at all. She is, in this sense, appropriately respons*ive* to reasons for regret, without having to be constantly respond*ing* to them.

I've now shown how the well-calibrated fitting utilitarian avoids three of the 'character-based objections' extant in the literature. Equipping the utilitarian agent with well-calibrated guiding dispositions helps to undermine claims that the fitting utilitarian psychology is so typically self-effacing (across a wide range of normal circumstances) as to warrant charges of intrinsic defectiveness. In the course of this defense, I have tried to employ a fairly conservative methodology: Beginning from assumptions about rationality that underpin the original objection, I have drawn out a conception of a *normally competent* fitting utilitarian agent, thus rebutting the charge that the fitting utilitarian is too incompetent to qualify as a truly morally fitting agent.

4.4 Act vs. Rule Consequentialist Agents

In light of my appeal to rules and dispositions, some readers may be puzzled by my labelling the resulting agent a fitting 'act utilitarian' agent, rather than a rule utilitarian one. To avoid any confusion on this front, let me wrap up by briefly characterizing what I take to be the two main differences between (fitting) act and rule utilitarian psychologies.

Firstly, while both make use of rules, they do so in very different ways. The act utiliarian adopts 'rules of thumb' for *instrumental* purposes, but their fundamental aim (in acting) makes no essential reference to rules: they just want to bring about the best possible outcome, and refraining from deliberation is one strategy they might employ, at appropriate times, as a means to this end. Rule Consequentialism, by contrast, builds reference to rules into its criterion of right action, and hence the corresponding 'fitting psychology' must likewise accord some fundamental, non-instrumental significance to rules. (This then opens them up to distinctively characterological objections of 'rule-worship'.)

A second, more straightforward difference is that they may employ rules with very different contents. I've suggested that a fitting act utilitarian could (whilst retaining their fitting character) only make use of 'well-calibrated' dispositions. But insofar as rule consequentialism appeals to rules that it would be good to internalize *for whatever reason*, they may well end up calling 'fitting' even dispositions that are merely extrinsically desirable. In other words, the fitting rule consequentialist agent would look much more like

the kind of 'sophisticated' agent described in §2. Insofar as we doubt that such a psychology *really is* morally or rationally fitting, this could provide the basis for a new argument against rule consequentialism — though not one that I have space to develop here.

References

Bales, R. E. 1971. "Act-utilitarianism: account of right-making characteristics or decision-making procedures?" *American Philosophical Quarterly* 8:257–65.

Chappell, Richard Yetter. forthcoming. "Value Receptacles." *Noûs* .

Gauthier, David. 1997. "Rationality and the Rational Aim." In Jonathan Dancy (ed.), *Reading Parfit*. Oxford: Blackwell.

Hooker, Brad. 2000. *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*. Oxford: Oxford University Press.

Kelly, Thomas. 2010. "Peer Disagreement and Higher Order Evidence." In Richard Feldman and Ted Warfield (eds.), *Disagreement*. Oxford: Oxford University Press.

Mackie, J. L. 1985. "Rights, Utility, and Universalization." In R. G. Frey (ed.), *Utility and Rights*. Oxford: Basil Blackwell.

Mason, Elinor. 1999. "Do Consequentialists Have One Thought Too Many?" *Ethical Theory and Moral Practice* 2:243–261.

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

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

—. 2011. *On What Matters*, volume 1. Oxford: Oxford University Press.

Pettit, Philip and Brennan, Geoffrey. 1986. "Restrictive consequentialism." *Australasian Journal of Philosophy* 64:438 – 455.

Railton, Peter. 1984. "Alienation, consequentialism, and the demands of morality." *Philosophy and Public Affairs* 13:134–171.

Slote, Michael and Pettit, Philip. 1984. "Satisficing Consequentialism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 58:139–176.

Stocker, Michael. 1989. *Plural and Conflicting Values*. Oxford: Oxford University Press.

Williams, Bernard. 1973. "A Critique of Utilitarianism." In J.J.C. Smart and Bernard Williams (eds.), *Utilitarianism: For and Against*. Cambridge University Press.

—. 1982. "Persons, Character and Morality." In *Moral Luck: Philosophical Papers, 1973-1980*. Cambridge University Press.