The Fitting and the Fortunate

Richard Yetter Chappell*
Princeton University

October 18, 2011

Abstract

Critics of consequentialism often object to *how a consequentialist agent would (allegedly) think*. They claim that the consequentialist agent is, in some sense, a *bad* character: cold and calculating, alienated from themselves and others, etc. Defenders of consequentialism typically dismiss such objections by citing the distinction between 'criteria of rightness' and 'decision procedures'. (Utility provides the criterion that determines the moral status of an act, but it's a further question whether agents ought to attempt to calculate utilities themselves. If it'd have bad results then consequentialists would recommend against it!) However, I argue that the strongest objection in this vicinity is not just that thinking like a consequentialist would have bad results, but that such a psychology would be morally *perverse*, in a sense that's incompatible with the psychology in question constituting a morally *accurate* way of thinking. I then assess several instances of this argument form, with particular attention to the 'value receptacle' objection.

*Thanks to Nate Gadd, Hrishikesh Joshi, Eden Lin, Brennan McDavid, Tim Mulgan, Philip Pettit, Peter Singer, Michael Smith, Helen Yetter Chappell, and an anonymous referee at *Nous*, for helpful comments.

Introduction

In order to assess the objection that the consequentialist agent would be of bad character, we must distinguish two very different kinds of normative evaluation: the 'fortunate', and the 'fitting'.[1] If a psychological trait is *fortunate*, or instrumentally valuable, then it will be recommended by the normative theory as something to aim *at*. Roughly, this is to say that it is desirable, or ought to be inculcated. For example, if caring intrinsically about your friends, or your stamp collection, is conducive to your own happiness, then egoistic hedonism would evaluate such concerns positively in this sense. But there's another important sense in which these are not properly *hedonistic* concerns. Reflecting this second kind of normative evaluation, we can ask whether the agent's psychology *fits with* or *embodies* the perspective of the normative theory — roughly, whether the agent is responsive to the reasons posited by the theory, in the sense that he desires what the theory says is desirable (one's own happiness, in case of hedonism), and has otherwise fully internalized the truth of the theory. This is to ask whether the agent's psychology is, in a sense, appropriate, rational, virtuous, or (as I will say) *fitting*.[2] Just as we make absolute assessments of rationality (i.e., rationality

[1] This mirrors the familiar distinction between state-given and object-given reasons for attitudes. See, e.g., Parfit (2011, 50).

[2] I assume that consequentialism aspires to claim the full normative authority of practical reason, or the all-things-considered 'ought'. But those who prefer to understand the view as *merely* concerning some rationally-optional 'morality' are free to read me as simply bracketing whatever other, non-moral considerations may intrude. My subsequent talk of 'rationality' should thus be understood as concerning just *what's rational from the moral point of view*.

"according to the true theory") in addition to assessments of what's rational according to any particular theory, so we can assess whether a psychology is virtuous or morally fitting *tout court*, or fits with the true moral theory, in addition to assessing whether it fits with any particular theory. Both "theory-relative" and "absolute" fittingness judgments will prove important in what follows.

What's fitting and what's fortunate may come apart, as in the 'paradox of hedonism': one is more likely to achieve hedonistically-fortunate results (i.e. happiness) if one does not possess the hedonistically-fitting mindset of adopting happiness as one's supreme goal. For another example: agent-neutral consequentialists might argue that things actually turn out better when people have more partial motivations (Jackson 1991), and so this is something we should want to encourage; but that doesn't change the fact that impartiality is really the more fitting or morally *accurate* perspective, according to their theory. It's just to say that some things are more important than being virtuous, or believing and fitting one's psychology to the truth.

With this distinction in hand, we can now identify two interpretations of the "bad character" objection to consequentialism. One claims that consequentialism is 'self-effacing' in the sense that it'd be *unfortunate* to possess the fitting consequentialist psychology. This is easily seen to be a poor objection. After all, it's always possible that events may conspire to punish those who are disposed to believe and act on the truth. Such circumstances do not cast doubt on a theory's claim to truth — again, it merely means that

it may be better, from a moral perspective, if we believed some false moral theory instead.[3]

However, when deontologists complain about the bad character of a committed consequentialist agent, there is a stronger objection that they may wish to present. This is the objection that the fitting consequentialist psychology is (contrary to the consequentialist's claims) *not actually morally fitting*. For example, they argue that the consequentialist agent is incapable of friendship or commitment to projects — but, they may add, this seems like an intrinsic defect: surely *genuine* virtue and rationality are not incompatible with these important goods. So, they conclude, the consequentialist's conception of rationality (virtue, fittingness) must be in error. Note that it's no response to this to say that consequentialism doesn't necessarily recommend that one try to become rational or morally fitting in this way. For the objection we're now considering is *not* merely that it would be inadvisable for agents to be ideally rational or fitting. That isn't the problem. Rather, the objection is that consequentialism implies something *false* about what the ideally rational or morally fitting mindset would *be* — whether we should seek to adopt it or not.

This objection is the real challenge, though it risks being obscured if we focus exclusively on evaluations of fortunateness. It's important to note that our moral theory also commits us to a conception of the morally or rationally

[3] A more sophisticated version of the self-effacingness objection is explored in my 'What's Fit for the Fallible' (ms).

fitting agent. This creates space for critics to question whether our conception of the fitting agent is really accurate. (Given the connection between value and fitting desire, such objections will prove equivalent to challenging the consequentialist's value theory *insofar* as the objections are concerned with fitting desire, but as we will see in §4, some characterological objections are broader in scope, and so can't be reduced to axiological objections.) To take up this challenge, we must either (i) bite the bullet and insist that what the deontologist identifies as intrinsic moral 'defects' are not really so, or else (ii) argue that, properly understood, the fitting consequentialist agent would not in fact possess the identified defects.

In this paper, I begin the latter task of rehabilitating our conception of the fitting consequentialist agent. In section 1, I identify and set aside a class of common objections that are only relevant to a peculiar subset of consequentialist views, so that we can turn our attention to characterological objections that purport to identify more general problems for consequentialism. Section 2 addresses various interpretations of the traditional 'value receptacle' objection. There I argue that it's a misconception to think that the fitting utilitarian agent would treat an individual's welfare as a fungible mere means to the end of aggregate welfare. Section 3 examines two influential objections from Bernard Williams, concerning whether a consequentialist agent would think either too much or too little about certain decisions. Finally, in section 4 I address the general worry that consequentialism embodies a perversely 'objective', mechanical attitude that would prevent the consequen-

tialist agent from properly relating to other persons. In all of these cases, I argue, the fitting consequentialist perspective does not in fact exhibit the perverse feature being attributed to it.

The objections thus surveyed are familiar ones. But we will find that it is illuminating to recast them in terms of fitting psychologies. There are three main reasons for this: One is that it allows for greater precision in specifying the objection and identifying its strongest form (this comes out most clearly in the discussion of replaceability in §2.2). Secondly, when the objection is thus clarified, it may suggest a natural new response. Finally, we may (in some cases) find ourselves with stonger intuitions about the moral appropriateness or perversity of concrete psychologies. So when long-running debates over abstract moral principles reach a dialectical stalemate, recasting them in a new light may allow more progress to be made.

1 Axiological Refinements

Before we begin, let me first set aside a class of objections that *aren't* relevant to the assessment of consequentialism as such, namely, those which trade on particular assumptions about *what* is of value (rather than how we should respond to whatever is of value). For example, we might consider it perverse to plug into Nozick (1974)'s experience machine, but if so, the lesson to take from this is not that consequentialism is wrong, but merely that the goods to be maximized include non-experiential goods. Similar lessons may

be drawn from 'Peeping Tom' objections: even if the victim never finds out about it, and hence experiences no pain or humiliation, we may still think that invasions of privacy *themselves* constitute a harm to the victim.

Axiological refinements also allow consequentialists to avoid the unpalatable 'mob rule' element of simple utilitarianism. It is entirely open to us to judge that sadistic pleasure, for example, has no positive value. So we need not think it appropriate for a sadistic majority to torture an individual, just because the quantity of pleasure thus obtained would factually outweigh the individual's pain. This larger quantity of pleasure may be of a nature that renders it *normatively* worthless.

Nor is this 'axiological' response to the problem merely an *ad hoc* move to save consequentialism. Rather, I think, anyone who considers this case a genuine problem to begin with should agree that the fundamental problem is one of disvalue rather than wrongdoing. We can see this by 'naturalizing' the scenario to replace the wrong action with a merely natural occurrence: Suppose that rather than a person torturing the minority individual, they were instead struck by lightning or some such. There is no action here, and so no wrongful action. Still, I have the strong intuition that this would be a bad thing to happen, no matter how much sadistic pleasure others might gain from their knowledge that this stigmatised individual suffered this harm. It would be very odd to think that sadistic torture was wrong, and yet that it would be a grand thing for the world if more stigmatised individuals suffered similar natural harms when accompanied by similarly greater net pleasure

for sadistic observers. Clearly the problem here, if there's a problem at all, is that increased sadistic pleasure does *not* strike us as a good outcome. But if that's right, then consequentialism will not tell us to bring about this outcome.

Not every appropriate axiological refinement can be identified via the above 'naturalization' test. This is because there may be intrinsic disvalue to particular acts or choices themselves: sometimes it is precisely the vicious or inconsiderate exercise of agency that contributes to the disvalue of an outcome. For example, imagine a doctor who secretly molests his patients while they're unconscious. We cannot really 'naturalize' this scenario without losing its key feature: If the doctor merely 'touches' the patient as a result of an involuntary muscular spasm, for example, then this no longer qualifies as a violation in quite the same way, and hence it does not seem nearly as bad.

Once we allow that consequentialists may build the intrinsic disvalue of vicious action into their axiology, we need a more sophisticated test to distinguish them from deontologists — and hence to distinguish genuinely anti-consequentialist intuitions from mere axiology-refining intuitions. At this point we may turn to agent- and time-neutrality. For while a consequentialist may be concerned to prevent vicious actions, she is ultimately no more concerned with her own actions than with other people's.[4] She will

[4] Here I assume that we are talking about *agent-neutral* consequentialism. Some agent-relative theories, e.g. egoism, are also plausibly consequentialist in nature, but then we can repeat the test using temporal neutrality rather than agent neutrality. Any theory that

thus consider it worthwhile to perform a single intrinsically bad action herself if this prevents multiple similarly bad actions from others. And it is much more difficult for the opponent of consequentialism to establish the intuitive repugnance of *this* than to poke holes in crude axiological theories like value hedonism.

In section 2, we will explore some non-axiological worries about 'replaceability' or 'value receptacles' that are more plausibly objections to consequentialism *as such*.[5] But before we do, let us first set aside two versions of the 'replaceability' worry that are merely axiological in nature.

1.1 Death and Replacement

One might worry about the fact that classical utilitarians attribute no significance to the ways in which experiences are packaged together into distinct lives, and hence only see death as bad insofar as it causes there to be fewer good experiences in future. We may worry that this does not do justice to the badness of death: most of us would not think it a good thing (all else equal) for someone to be struck down in the prime of life and replaced with a marginally happier substitute. The premature death of an individual is bad in a way that goes beyond the mere failure to create future goods. Death is

builds in time-relative 'side constraints' that should not be violated now even to prevent more such violations in future, sounds fundamentally deontological to me. Even if one could model the theory using time-relative values to be maximized, as in Louise (2004), this seems to distort the motivation for the theory. See my first chapter, Fitting Attitudes for Consequentialists, for more detail.

[5] We will see that there is a subtle sense in which even the later objections qualify as 'axiological', but they will concern the general *structure* of our value theory rather than its particular *contents*.

not equivalent, as this view would have it, to the failure to create life.

But we can accommodate this intuition without abandoning consequentialism. We merely need to refine our axiology so as to properly capture the disvalue of death. Here's one possibility: Besides preventing the creation of future goods, death is also positively disvaluable insofar as it involves the interruption and thwarting of important life plans, projects, and goals.[6] If such thwarting has sufficient disvalue, it could well outweigh the slight increase in hedonic value obtained in the replacement scenario. Consequentialists are thus fully able to attribute significance to the packaging of experiences into lives, and to acknowledges the positive disvalue of death — they just need the right theory of value.

1.2 Imprecise Values

One might also object to (an implication of) the traditional consequentialist practice of assigning exact numerical values to things.[7] Suppose we begin with two people, neither of whom has a more valuable life than the other, and you can save only one. It doesn't seem that mildly "sweetening" one of the options, with a dollar bill or the like, should break the tie or make the choice any easier or less arbitrary. Consequentialists may accommodate this phenomenon of *resistence to sweetening* by — once again — appropriately

[6] Importantly, this account of the positive badness of death avoids the opposite mistake of attributing *constant* and *unconditional* disvalue to death. There may be circumstances in which death is an unmitigated blessing, after all. Instead whether — and to what extent — death constitutes a positive harm will depend on the situation, i.e. what important life projects it cuts short.

[7] Thanks to Daniel Greco for bringing this to my attention.

complicating their value theory. Rather than holding the two lives to be precisely equal in value, they must be merely *roughly* equal (Parfit 1984, 431), or 'on a par' (Chang 2002), such that sweetening one option does not necessarily make it of greater total value than the other (despite being better than it was prior to sweetening).

While there is some intuitive support for the thought that resistence to sweetening is often appropriate, I don't think that it would be at all *immoral* to insist on precise values.[8] As we will see, it's a mistake to think that treating people's lives as *comparable* in value entails treating them as *fungible* or interchangeable in the way that we treat money, for example, as being. I might be genuinely torn between two distinct but equal intrinsic values, recognizing the separate force of each, even as my decision hangs in the balance such that the slightest inducement to either side would sway my decision. The sensitivity of my decision to further incentives does not in any way imply a failure to appreciate the distinct and irreplaceable conflicting values in play. So the separateness (or non-fungibility) of values cannot be understood merely as a matter of their being not precisely comparable. We need a better account. In the following sections — most notably §2.2 — I advance a positive account of what it really takes to disrespect a person by treating them as fungible, and how consequentialists can avoid this fate.

[8] While perhaps a "moral error" in some abstract sense, it is not *disrespectful* of another's person in the sense discussed in later sections.

2 Value Receptacles

One traditional objection to consequentialism — expounded, for example, in (Regan 2004) — is that it constitutes a perspective from which individuals are seen as mere 'receptacles' or repositories for whatever happens to be of value: let's stick with happiness, for simplicity. The general worry is that a genuinely consequentialist agent would fail to recognize other individuals as valuable ends in themselves; instead, the objection goes, individuals are seen as merely *instrumental* to the end of realizing happiness (or value more broadly) in the world. This objection may be refined in a couple of different ways.[9] Let us consider them in turn.

2.1 Incidental Interests

It's widely agreed that we have reasons to help other people. But we may ask about the deeper structure of these reasons: *why* do we have this reason? On whose behalf does this reason exert its normative force or make claims on us? The commonsense answer is that these normative reasons speak on the behalf of the individuals who need our help. It is *for their sake* that we have reason to relieve their suffering. This much seems clear.

Yet utilitarians might be thought to deny this datum. As Singer (1993, 121) puts it, "The total version of utilitarianism regards sentient beings as valuable only in so far as they make possible the existence of intrinsically valuable experiences like pleasure." There is no mention here of the interests

[9] Thanks to Pablo Stafforini for helpful discussion on this point.

of the beings experiencing these pleasures. If the utilitarian's theory simply tells her to maximize net happiness, it may seem natural to reconstruct the fitting utilitarian's thought-process as follows: *Bob is in agony. My goal is to maximize utility, i.e., the balance of pleasure over pain. There is some agony (namely, Bob's) that I am in a position to relieve. Doing so would serve my goal. So I will act to relieve Bob's suffering.* But now note that the interests of *Bob himself* seem to have dropped out of the picture for our imagined utilitarian agent. She is merely concerned to minimize pain and suffering. The fact that doing so is *good for Bob* (or anyone else) is not a relevant consideration to her way of thinking, or so we might imagine. Helping people is incidental, a mere side-effect to her real goal of patterning the universe with a particular class of experiences. Call this view *Utility Fundamentalism*.

By taking the value of pleasure (and disvalue of pain) as fundamental, and not to be explained in terms of their value *for* individuals, Utility Fundamentalism seems objectionably fetishistic. It treats individuals as intrinsically valueless 'receptacles', of moral interest only insofar as they provide a space or habitat for what (supposedly) really matters: the brute promotion of pleasure over pain. This moral perspective strikes us, I think rightly, as perverse.

If this is how we are to understand the 'value receptacle' objection, then utilitarians (and consequentialists more broadly) may escape it simply by rejecting Utility Fundamentalism. After all, there is a very natural alterna-

tive account, according to which pleasure (say) is good precisely *because* it is good *for* the individual who experiences it, and suffering is bad because it is bad for the suffering individual (Wilson 2006). On this view — call it *Welfarism* — the interests of individuals play an essential explanatory role in our value theory. When the welfarist utilitarian relieves Bob's suffering, the fact that this benefits Bob is not merely incidental to her reason for acting. It is, on the contrary, the source or ground of her reason. She has reason to relieve suffering precisely because this is good for someone.

We may demonstrate the difference between these two views by way of a fanciful counterfactual: If the welfarist utilitarian became convinced that some pain was, for some reason, intrinsically good for Bob, she would no longer take herself to have non-instrumental reason to rid Bob of it. The utility fundamentalist, by contrast, has a fixed goal that makes no mention of the interests of individuals *as such*. She cares about experiences, not experienc*ers*. So even if she too believed pain to be good for Bob rather than bad for him, this would be of no intrinsic interest to her: she just wants to minimize pain, no matter whether this helps or harms the individuals experiencing the pain in question.

We thus see that only the utility fundamentalist is liable to the 'value receptacle' objection, understood as the failure to recognize that happiness (or whatever) is good just because it's good *for* individuals. This fetishistic perspective is by no means endemic to consequentialism. Indeed, it is entirely natural for consequentialists to instead take the welfarist route of specifying

that happiness is good precisely *because* it's good for the individual who experiences it. Our current interpretation of the value receptacle objection is then simply inapplicable to this welfarist form of consequentialism.

2.2 Are Persons Replaceable?

Even given an appropriately welfare-based explanation of why happiness matters, there remains a second interpretation of the 'value receptacle' objection that might be leveled against the utilitarian. The remaining objection is that utilitarians treat particular individuals not as ends in themselves, but merely as fungible or replaceable means to the end of promoting *aggregate* welfare.

This objection has been formulated in several different ways. Rawls (1999, 24) famously objected that "Utilitarianism does not take seriously the distinction between persons." Singer (1993, 121) writes, "It is as if sentient beings are receptacles of something valuable and it does not matter if a receptacle gets broken, so long as there is another receptacle to which the contents can be transferred without any getting spilt." The common thought here is that there's an important sense in which utilitarianism fails to treat us *as* individuals. It takes our interests into account, perhaps even *as* interests, but not in a way that appreciates the normative *distinctness* of my interests and yours. We are all melded together, into a kind of unstructured, undifferentiated welfare soup.

These formulations are evocative, but imprecise. I think we get a firmer grip on the objection by formulating it in explicitly psychological terms.

The fitting utilitarian agent would (allegedly) have but a single ultimate desire: to maximize aggregate welfare. They thus see different individuals as interchangeable. It makes no difference, to such an agent, which of several people is helped (or indeed whether one person is helped a lot or several people each helped a little), so long as the impact on aggregate welfare would be the same in either case.

To bring out why this is so objectionable, note that fungibility is, in general, the mark of the instrumental. Money is fungible precisely because we do not value the possession of *particular* bills: replacing two tens with a twenty would serve my ends just as well. For another example, if my sole ultimate desire is to slake my thirst, then I will be indifferent between two equally effective means to satisfying this goal. If someone switches my glass of water for another that's qualitatively identical, this is not a change that's normatively significant to me. I do not desire *that* glass in particular, so it may just as well be replaced by any other that would do the job. On the other hand, if I *had* (bizarrely) desired the original glass in its particularity, then the substitution would be of significance to me: it would thwart one of my non-instrumental desires.

This connection between fungibility and merely instrumental valuation explains why the above objection to utilitarianism seems so forceful. It seems perverse to treat individuals as replaceable or fungible, because such treatment constitutes a failure to intrinsically value individuals in their particularity. The correct moral theory, we feel, must attribute intrinsic value to

particular individuals and not just to the general welfare.

How is a theory to satisfy this requirement? Again we can clarify the matter by reference to fitting psychologies. We have seen that it's morally perverse for an agent to be *indifferent* between options that equally benefit distinct people, for that is to disrespect the individuals by treating them as fungible means to the aggregate welfare. But of course we do not want to favour either person over the other, since such bias would constitute disrespect for the person whose equal benefit we counted for less. Instead, I propose, the fitting response to a tradeoff between two distinct but equally weighty values is to feel *ambivalent* about the choice. There are distinct reasons pulling you in either direction, corresponding to the distinct values served by either choice. But these reasons are equally weighty, so the agent is *torn* rather than pulled without resistance towards one choice over the other.

This is a distinction we should want our theories to be able to make. Whatever substantive disputes we may have about what is of value, we should all acknowledge the formal difference between (i) a pair of options serving distinct but equally weighty final values, and (ii) a pair of options that serve literally one and the same value. For example, assuming that token artworks have intrinsic value, a choice between saving a great painting or an equally great sculpture is importantly different from a choice between saving the same painting in either of two different (but equally effective) ways. In the latter case, the two options are seen to serve the same token value in virtue of saving the same token artwork. Other cases of this may be more subtle,

as even two distinct concrete objects may serve as vessels for one and the same token value. An intuitive example of this is pleasure: I'm completely indifferent between the prospects of a massage for my left foot or my right, assuming that either would be similarly pleasant.[10] I take this to suggest that left-foot-pleasure and right-foot-pleasure are not distinct final values, the way that the painting and the sculpture (or my welfare and your welfare) are. Instead, it seems, I ultimately value pleasures of a certain qualitative kind *in the aggregate*, and particular instances of such pleasures are thus, in an important sense, of merely 'instrumental' value to me. Of course this is not to say that they are causally instrumental to some downstream effect. We may instead call it *constitutive instrumentality*, as each token of pleasure is a constitutive, rather than causal, means to the end of aggregate pleasure.

With this understanding in hand, we may now characterize the replaceability objection as alleging that fitting consequentialists must likewise treat *individual persons* as constitutive means to the aggregate welfare, rather than as distinct ends in themselves. Given that individual persons have final value, such instrumental treatment constitutes a distinctive kind of *disrespect* or failure to respond appropriately to the value that persons have in themselves.

The reader should now have an intuitive grasp of the distinction between (equally-weighty) distinct final values and (equally effective) mere means to a single final value. I've suggested that one way this distinction might play

[10] Thanks to an anonymous referee for prompting me to discuss this case.

out is that in the second case the two options are perfect substitutes, and hence the fitting attitude for an agent to take towards them is indifference. In the former case, by contrast, the two options are not *substitutes*; they serve different ends, albeit equally worthy ones. This naturally suggests that the fitting attitude to take is ambivalence, rather than indifference.

Another way to support this conclusion is via the idea that it's fitting to intrinsically desire *each* intrinsic good, with strength proportional to the magnitude of the object's value. If, and only if, a pair of options serve distinct intrinsic values, will the two options differentially satisfy the intrinsic desires of the morally fitting agent (and hence strike her as significantly distinct). Insofar as the agent has conflicting desires, we can say that she manifests ambivalence rather than indifference over the options.

We are now in a position to evaluate the objection that utilitarianism treats people, and their interests, as fungible. This is, as we have seen, equivalent to interpreting utilitarianism as the view that only one token thing, namely aggregate welfare, has intrinsic value. Call this view *monistic utilitarianism*. This view really does neglect the separateness of persons, for it attributes intrinsic value merely to the whole, and not to each of us in our particularity. As a consequence, the fitting monistic utilitarian has but a single desire — to maximize welfare — and treats our individual interests and concerns as mere (constitutive) means to the satisfaction of this more global goal. This is, I agree, morally perverse.

But there is no reason why utilitarianism must take this monistic form.

There is a very natural alternative view, call it *pluralistic utilitarianism*, on which *each* particular person's interests are (separately) accorded final value.[11] There is not just one thing, the global happiness, that is good. Instead, there is my happiness, your happiness, Bob's, and Sally's, which are all equally weighty but nonetheless distinct intrinsic goods. What this means is that the morally fitting agent should have a corresponding plurality of non-instrumental desires: for my welfare, yours, Bob's, and Sally's. Tradeoffs between us may be made, but they are acknowledged as genuine tradeoffs: though a benefit to one may outweigh a smaller harm to another, this does not *cancel* it. The harm remains regrettable, for that person's sake, even if we ultimately have most reason to accept it for the sake of more greatly benefitting another.

Contrast this with the case of money: If you have to invest $5 to earn $10, there is nothing to regret. The $5 is a "cost" merely in the sense that it would have been *even better* if you could have attained the $10 payoff without having to pay the $5. But given that this is not an option, there is nothing regrettable about the deal *as a whole*, the way that there is something regrettable about benefitting one person greatly at lesser cost to another. We can explain the difference, in the cash case, as a matter of both sums of money being mere components or constituents of the single token value,

[11] The view may still be monistic in the sense that there's just one *type* of thing that's good (cf. Hurka 1996). But the crucial point for present purposes is that there are a plurality of *token* final values. The separateness of persons merely requires that we each be valued separately. There's nothing obviously objectionable about it turning out that we are valuable in the same kind of way.

or desirable end, of aggregate wealth. This is very different from how the pluralistic utilitarian conceives of welfare tradeoffs between distinct persons.

We thus find that (pluralistic) utilitarianism is well able to reflect the normative separateness of persons, and to avoid treating people as fungible, replaceable receptacles of value. This is, if correct, an important result: It's commonly thought that the utilitarian's willingness to weigh harms to one person against benefits to another essentially involves treating the one as a "mere means". But my above analysis suggests that this traditional thought is simply confused. One may have thoroughly *non-instrumental* desires for each of two distinct instrinsic goods, and make reluctant tradeoffs between them in a way that is importantly different in kind from the tradeoffs one makes with fungible goods like money. The mere willingness to balance conflicting values is not itself constitutive of instrumental or fungible treatment. Critics may still insist that utilitarianism is just *extensionally incorrect* in its prescriptions for morally right action, but those wanting to make stronger claims about 'value receptacles' need to back up their claims with a rival account of instrumental valuation — as such rhetoric is seen to be baseless if the present account of instrumental valuation is correct.

An interesting implication of my account is that we may find that we actually treat our interests-at-a-time as fungible.[12] While we might initially have assumed that our momentary interests have final value, we may find on reflection that we consider our interests across time, unlike interests across

[12] Thanks to an anonymous referee for bringing this to my attention.

people, to be properly fungible. As in the case of fungible pleasures, this view can easily be incorporated into my framework by positing that individuals' interests-at-times are mere constitutive means to the final good of their timeless welfare. Alternatively, you might opt for the view that it's fitting to consider tradeoffs between timeslices to be just as emotionally fraught as tradeoffs between persons, and so assign final value to each momentary self individually. For purposes of this paper, I can remain neutral on this question of whether to attribute final value to momentary welfare, or only to timeless welfare.

2.3 Objections

I have argued that the fitting utilitarian could respect the distinctions between persons by separately desiring the good of *each* person's welfare, rather than having a single, totalizing desire for the aggregate good. But a difficulty arises when we consider goods that the agent is unaware of. Consider some particular unknown person, Harry. Our utilitarian cannot have a particularized desire for Harry's welfare, since she cannot even refer to Harry in particular. But her values must extend to others somehow: It's not as though she'd accept an offer to improve the welfare of her neighbour Bob at greater cost to some unknown other. So it seems that we need something like a generic desire for aggregate welfare to step in and fill the gap. (To avoid double-counting, we'd probably need to exclude Bob — and any others for whom the agent already has a particularized concern — from the remaining

aggregate.)

Is this a problem? Perhaps not. It doesn't seem so objectionable to treat people you've never even heard of as faceless members of the aggregate. How could they be *other* than faceless and generic to you? Moreover, the agent's attitude here is not merely instrumental. It's not as though our utilitarian thinks that unknown people fundamentally matter only in respect of their being members of the unknown aggregate. Rather, her concern for unknown people's aggregate welfare is a stop-gap measure that reflects, in the only way possible, her appreciation of the fact that each of those individual unknown persons fundamentally matters in their own right. She knows that, if she knew more, she would form particularized desires for the welfare of each; but in the absence of the requisite identifying information, the best she can do to respect these unknown values is to fall back on the generic desire for aggregate welfare, as a kind of placeholder.

So far, so good. But what about *merely possible* future persons? (Compare Parfit (1984)'s 'Non-Identity Problem'.) Here the placeholder strategy seems dubious. Before, we were holding the place for the particularized desires we would have if fully informed — and it seems reasonable for an agent to deferentially ascribe normative authority to her fully-informed desires. But in case of merely possible persons, the barrier to particularized reference is metaphysical, not merely epistemic: There *is* no such particular person to refer to. The most we can appeal to is the *counterfactual* desire that we (ideally) would have had if someone else had existed. We would

have formed a particular desire for that someone's welfare. But so what? As things stand, there is no such person, and hence no valuable entity for us to respect as best we can. We cannot have a 'second-personal' reason (Darwall 2006), grounded in the normative authority of the non-existent individual himself, to take his possible welfare into account. Our concern must instead be, in a sense, 'impersonal'.

Even so, this needn't return us to any single, totalizing desire that the world be thus-and-so. We may instead have distinct desires for each possible generic good. We still need to distinguish between indifferent and ambivalent pairs of prospects, after all. For example, I may desire both that Anne have a happy child rather than none at all, and that Beth have a happy child rather than none at all. Perhaps I cannot coherently desire these things *for* the sakes of the respective children (especially if they never actually exist), but I can desire them — for the sake of the world, perhaps. And in so doing, I recognize that the prospective persons are not fungible, in the following sense: Despite being of equal value, there is a morally relevant difference between a world where only Anne has a child, and a world where only Beth has a child. The comparison calls for ambivalence, rather than indifference, since they serve distinct (though equally weighty) ends or ideal desires. If Anne's child would have a better life, then I could prefer that she be the one to come into existence, even while I regret the absence of Beth's possible child, whose life would have been (distinctly) intrinsically valuable in its own right.

The objector might respond by suggesting that it's only because of the differential impact on the existing individuals Anne and Beth that we see a significant difference here. If we imagine some more thoroughly generic question — say, whether the 100th child born in the year 2500 is a boy or a girl — indifference may seem the only appropriate response. I'm not sure about this, as our lack of response may just be due to our contingent failure to really vividly appreciate what a significant intrinsic difference the identity of each individual makes. But even if the critic is right here, it's not clear that this is any objection to consequentialism in particular. If it turns out that distant future people cannot *but* be thought of as fungible, in the noted sense, then this limitation will presumably apply to all moral theories.

So I think the objection ultimately fails. Even in the toughest case — that of merely possible persons, who cannot be the ultimate ground of our concern for their welfare — consequentialists can plausibly still desire each good separately, and hence refrain from treating people as fungible. And even if it turns out that I'm mistaken about this, and in fact merely possible persons *are* fungible, then that is no fault of consequentialism. It would instead be a constraint that any moral theory must work within. So the remaining challenge for a theory would just be to ensure that it doesn't inappropriately extend the domain of the fungible to include actually existing persons. Consequentialism can, as I have shown, meet this challenge.

3 Thoughts Too Many, or Too Few

Bernard Williams has objected that consequentialists would find some choices more obvious than they should be, and that there are other actions that they would inappropriately pause over. As an example of the first, consider *Jim and the Indians* (Williams 1973): Captain Pedro will kill twenty innocent locals, unless Jim elects to kill one of them. What should Jim do? Assuming that all else is equal, and Jim knows it, then agent-neutral consequentialism implies that he should kill the one. The mere fact that it is *he*, rather than Pedro, who does the killing here is irrelevant so far as choosing between the two options is concerned. I argue elsewhere[13] that rational heuristics for non-ideal agents might complicate the matter — as may our normative uncertainty, insofar as the truth of consequentialism is itself unobvious (Lenman 2004) — but in this paper I'm focusing on the ideal case. And in the ideal case, where Jim has a full grasp of the situation and no cognitive limitations, I think the consequentialist should simply insist that the choice *would* (or should) be obvious.

Williams (1982) puts forward the trickier charge that consequentialist agents would have "one thought too many". When such an agent jumps into the water to save his drowning wife, Williams writes, "it might have been hoped... that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this

[13] See 'What's Fit for the Fallible', the third chapter of my dissertation.

kind it is permissible to save one's wife." (p.18)

The objection here is presumably not that one *blindly* ought to save one's wife, no matter the consequences for others. That would not be plausible. Rather, I take it, the challenge is to clarify the proper *role* that moral considerations play in one's cognitive economy. In particular, we may need to distinguish between *reasons* and background *conditions* on the applicability of a reason (compare, e.g., Schroeder 2007, chapter 2). In the case at hand, we may think that the agent should be motivated *just* by a concern for his wife, though his acting on this motivation is contingent upon the fact that he isn't causing greater harm by doing so. I don't see any barrier to consequentialism taking this distinction on board. So just because the consequentialist agent *wouldn't* have saved his wife if it hadn't been permissible for him to do so, it doesn't follow that the fact of its permissibility entered into the content of his 'motivating thought'. The condition may have a merely *virtual* presence in the agent, in Pettit (1994)'s sense that the violation of the condition would *trigger* the agent's attention and reconsideration; but in the absence of this trigger it doesn't figure in his thoughts at all.

A second way to interpret the objection here is that it would seem a kind of moral fetishism for an agent to be motivated by considerations of moral duty *as such*, i.e. under that abstract description. But, as I argued in section 2.2, there is no good reason to interpret the fitting consequentialist agent as having such abstract motivations. Rather, we should understand them as having intrinsic desires corresponding to each concrete intrinsic good — such

as the welfare of any given person. So understood, the consequentialist is moved directly by his concern for his wife, unmediated by any more abstract description of 'duty' or 'the general good', even while he remains sensitive to the interests of concrete others in a way that ensures he is not blind to duty or the general good.

A third worry suggested by this case is that the consequentialist is responding to his wife no differently from how he would respond to anyone else. A few things can be said about this. One is that, lacking omniscience, we are aware of the virtues of our loved ones in a way that we are not aware of strangers. So even an agent with a standing disposition to value all persons equally might well find the value and needs of people he is more familiar with to be more salient, and hence more emotionally motivating. More importantly, when we consider the ideal case, impartial consequentialists can offer an attractive response: though it may be psychologically impossible for us humans, ideally an agent really ought to care about *all* persons as intensely as we care about our loved ones. Finally, I should note that if one remains unattracted to impartiality as a moral ideal, there is always the option of adopting an agent-relative axiology, weighting the welfare of one's loved ones more heavily than that of mere strangers.

4 The Participant Attitude

Strawson (1962) famously distinguished two broad stances or attitudes we

might take towards another person. One is the attitude of "involvement or participation in a human relationship", which might naturally give rise to such personal reactive attitudes as gratitude, resentment, and so on. The other, which Strawson calls the 'objective' attitude, he describes in section 4 of his paper as follows:

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; [...] to be managed or handled or cured or trained; perhaps simply to be avoided [...] If your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him.

We might summarize the distinction by saying that the objective attitude involves seeing the other as an object, or a force of nature, to be understood and perhaps manipulated towards desired ends, whereas the participant attitude involves seeing the other as a fellow agent — a co-participant, perhaps, in the collective project of living well.

This distinction may help us to elucidate the unease that many feel about consequentialism. For they may have a sense that consequentialism represents a distressingly mechanical or 'objective' view of the world, where one's fellow agents — and even one's own future self — are seen primarily as causal levers rather than as rational beings. Utilitarian discussions of punishment, for example, are centrally concerned with the mechanical question of how to

bring about better outcomes. Consequentialists will generally have a lot to say about when it is or isn't useful to blame people, but tend to put aside questions of blame-*worthiness* as empty or misguided.

This exclusive focus on outcomes might seem to neglect some of the most important questions, for in our everyday relationships we are often more interested in what emotional responses would be *fitting* than in what responses would be *best*. It seems doubtful whether an agent who failed to share these concerns would even be capable of the sorts of genuine relationships we ordinarily (and, I think, rightly) prize. There's plausibly something *defective* (and not merely extrinsically unfortunate) about an agent who is constitutionally incapable of relating to others in this way. So it is of the utmost importance for moral theorists to uncover whether the fitting consequentialist agent really *would* forsake the participant attitude in this way. For if so, this would provide grounds for a decisive objection to consequentialism: its conception of the 'fitting' moral agent would not be genuinely morally fitting at all.

I will address this challenge in two parts, first examining how consequentialist agents would relate to themselves, and second, how they might relate to others.

4.1 How we relate to ourselves

Some may worry that the consequentialist agent takes an objectionably 'objective' or mechanistic attitude, not just towards others, but even towards

their future selves. Suppose that I am currently faced with a choice between A and B, and that tommorrow I will be offered a choice between two other options, C and D. Let us suppose that the value ordering of the possible outcomes is as follows: A & C > B & D > B & C > A & D. That is, my choosing A today could lead to either the best or the worst outcome, depending on whether I choose C or D tomorrow. Further suppose that I know, from past experience of similar choices, that I'm much more likely to choose D than C tomorrow, whatever I choose today. How should this fact affect my present decision-making?

According to 'Possibilists', it shouldn't affect my decision at all. It's within my power to choose A now, and C later, and so that's precisely what I should do. 'Actualists', by contrast, may admit that while A & C is the best available act-*sequence*, the decision facing me at the current moment is just what to choose *now*, and in light of the terrible expected value of now choosing A (since it will most likely be followed by my later choosing D), I should instead pick the safer option B (Jackson and Pargetter 1986).

Actualism is, I believe, the view that is more true to the spirit of consequentialism: We should act in the way that would maximize expected value, all things considered — and one such thing to consider is our disposition to act wrongly in future. But one might object that this is to take an inappropriately alienated perspective towards one's future self. It is to treat one's future decisions as beyond one's (current) control, which may seem wrongheaded. This is how one might find it objectionable how the consequentialist

agent relates to him or herself.

I think this objection is misguided. It is a factual question whether there is anything that the agent can do now (including mental activities like steeling her will) to ensure that she acts rightly in future. If there is something (A*) that the agent can now do to ensure this ideal outcome, then the Actualist will obviously join the Possibilist in recommending this option. On the other hand, if it's really *true* that whatever intentions she forms now won't make any (or enough) difference to her future actions, then it cannot be inappropriate for the agent to act in light of this truth. Indeed, it would seem plainly unwise for her *not* to! So I think our intuitive resistance to Actualism actually stems from imaginative resistance to the set-up: We tend to assume that, in choosing A, the agent *could* (concurrently) do something to secure her future compliance with the ideal plan. But we can only distinguish Actualist and Possibilist recommendations if we imagine a case where this ordinary assumption fails to hold. In such abnormal cases, I think the Actualist's prescribed action is the sensible choice for the agent. It really is appropriate to consider your future self as beyond your current control in those odd cases where this is true. On the other hand, Actualist Consequentialism does not recommend that one *always* adopt this alienated self-perspective. On the contrary, in all those ordinary cases where you can currently form intentions that will secure future right actions, that's precisely what you should now do![14]

[14] Doug Portmore has further developed this basic idea, which he calls 'Securitism', in

4.2 How we relate to others

We've seen that consequentialist agents would intimately identify with their future selves, when appropriate, by deliberating now about future actions. But this *deliberative* identification is obviously inapplicable to *inter*-personal relations. (We may decide how to act in future, but we cannot directly decide how others will act. At most we can decide what instructions, requests, or threats we will send their way.) So, in order to assess the force of the objection that consequentialists would not appropriately relate to others, we must first get clearer on what kind of interpersonal relation is appropriate.

In the interpersonal case, the relevant difference does not show up so clearly in what decisions we would make, but rather in what we would care about, and how we would respond emotionally to others' apparent attitudes towards us. As previously described, the participant attitude involves respecting the other person *as* a person, in a way that involves caring about whether they likewise respect us. When we adopt the participant attitude towards someone, we may respond with gratitude when they show a surfeit of good will towards us, and with resentment when they are inconsiderate or malicious. If we adopt the objective attitude, by contrast, we see the other from a much more detached and impersonal perspective, and are no more emotionally vulnerable to them than we would be to a valuable vase or a gushing waterfall. We may still want to protect these valued objects — the objection here is not that the consequentialist fails to value people. Rather,

chapter six of his (Forthcoming).

the objection is that one who exclusively adopts the objective attitude is failing to respond to other people in the appropriate *way*. Though we may in some sense be valuable objects, that is not *all* that we are. We are also rational *subjects*, a fact which calls for recognition from our fellow agents.

I'm not convinced that there is really any incompatibility between consequentialism and the participant attitude. Consequentialism as a theory is focused on a different question — namely, how we should act.[15] But then the present objection merely suggests that consequentialism is incomplete, and we should seek to supplement this theory of fitting action with a theory of fitting reactive attitudes (gratitude, resentment, etc.). That is, a fitting consequentialist *may* fail to take up the participant attitude, but she need not so fail, and indeed if *fully* normatively fitting she would take it up when appropriate. Note that this is no more an objection to consequentialism within its domain than is the need to supplement it with an epistemological theory of fitting belief. There's more to normativity than just actions, and the right norms for other attitudes may differ from the consequentialist norms that are fitting for action. Consequentialists should, I believe, accept this fact about the limited scope of their theory.

One might press the objection by pointing out the practical incompatibility of the participant and objective attitudes. Insofar as the objective at-

[15] For a defence of this conception of consequentialism, contra Pettit and Smith (2000), see my first chapter, 'Fitting Attitudes for Consequentialists'. Let me also clarify that the consequentialist's axiology will have implications for what it's fitting to desire, but this still leaves open what the correct norms are for various emotional states, etc.

titude — attending to the causal structure of the world, including its human inhabitants — better prepares one for consequentialist decision-making, and adopting this attitude is in practice incompatible with simultaneously adopting the participant attitude, then doesn't consequentialist decision-making effectively *preclude* one's sharing in the participant attitude? Here I find it illuminating to consider Bennett (ms)'s tentative suggestion for how to understand the practical tension between the participant and objective attitudes:

> Reactive attitudes essentially prepare for personal interaction of a certain kind, while the objective attitude prepares for inquiry, and these two sorts of activity are somehow incompatible. If that is right, the two sorts of attitude are derivatively in conflict, like simultaneously readying oneself for a sexual encounter and for giving an after-dinner speech.

This account of the role of the two kinds of attitude seems plausible: There's something very relationship-oriented about the participant attitude, whereas the objective attitude seems forced on us when our task is to make an accurate prediction or accounting of things. But if we accept this, it allows us to dissolve the present objection, as there's no reason to think that agents (including consequentialist agents) must be incessantly inquiring. Admittedly, we often consider tricky cases of the sort where inquiry into the likely consequences of various possible decisions is indeed called for. So it is un-

derstandable that some would come to associate consequentialism with the objective attitude. But ordinary life does not call for constant inquiry,[16] so there is plenty of space left in the agent's life for the participant attitude to play its role.

Conclusion

In this paper, I have distinguished two general forms that a character-based objection to consequentialism might take, mirroring the distinct normative assessments of what's *fortunate* and what's *fitting*. I suggested that objections from fittingness are more powerful, and that these objections are not adequately addressed by the standard consequentialist strategy of distinguishing criteria of rightness from decision procedures. I then took some initial steps towards formulating an adequate response to the fittingness objections against consequentialism. In particular, I showed that the consequentialist perspective does not involve seeing individuals as fungible means to the general good, that the consequentialist agent need not have 'one thought too many', and that they are not barred from adopting the 'participant attitude' that is a necessary prerequisite for interpersonal relationships. I do not pretend to have addressed *every* important character-based objection to consequentialism within these few pages, but just to have made a start that needs to be further built upon in future work.

[16] This is related to the problem of 'defective deliberateness' that I further discuss in my third dissertation chapter, 'What's Fit for the Fallible'.

References

Bennett, Jonathan. ms. "Accountability." http://www.earlymoderntexts.com/jfb/accounta.pdf.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112:659–688.

Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Hare, Caspar. 2010. "Take the Sugar." *Analysis* 70:237–247.

Hurka, Thomas. 1996. "Monism, Pluralism, and Rational Regret." *Ethics* 106:555–575.

Jackson, Frank. 1991. "Decision-theoretic consequentialism and the nearest and dearest objection." *Ethics* 101:461–482.

Jackson, Frank and Pargetter, Robert. 1986. "Oughts, options, and actualism." *Philosophical Review* 95:233–255.

Lenman, James. 2004. "Utilitarianism and Obviousness." *Utilitas* 16:322–325.

Louise, Jennie. 2004. "Relativity of value and the consequentialist umbrella." *Philosophical Quarterly* 54:518–536.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

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

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

Pettit, Philip. 1994. "Consequentialism and moral psychology." *International Journal of Philosophical Studies* 2:1 – 17.

Pettit, Philip and Smith, Michael. 2000. "Global Consequentialism." In Brad Hooker, Elinor Mason, and Dale Miller (eds.), *Morality, Rules, and Consequences*. Edinburgh: Edinburgh University Press.

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

Rawls, John. 1999. *A theory of justice*. Belknap Press of Harvard University Press, revised edition. ISBN 9780674000780.

Regan, Tom. 2004. *The Case for Animal Rights*. Berkeley: University of California Press.

Schroeder, Mark. 2007. *Slaves of the Passions*. Oxford: Oxford University Press.

Singer, Peter. 1993. *Practical Ethics*. New York: Cambridge University Press, second edition.

Strawson, Peter. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48:1–25.

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.

Wilson, Scott. 2006. "Respect for Utilitarianism: A Response to Regan's 'Receptacles of Value' Objection." *Proceedings of the Ohio Philosophical Association* 3.