**THE BEST EXPRESSION OF WELFARISM**[[1]](#footnote-1)

Christian Coons

The term ‘welfarism’ is used in a variety of ways in moral philosophy, political theory, and economics. In this paper, following L. W. Sumner, it refers to a view about the foundations of morality—the view that “nothing but welfare matters, basically or ultimately, for ethics” (Sumner 1996: 184).[[2]](#footnote-2) Again following Sumner, welfarists believe:

*Axiological Monism*: The list of foundational goods contains a single item: well-being. Only states of individual well-being are intrinsically good.

*The Priority of the Good*: Facts about which states are intrinsically good explain which duties and moral reasons we have. The point of ethics is to bring about good (or better) states of affairs.

*Agent-Neutrality*: The foundational ethical values are agent-neutral—there is a reason for each agent to promote it, or aim to bring it about. (Sumner 1996: 184–5)

I do not defend welfarism, though I briefly rehearse some of its attractions. Instead, my concern is which specific normative theory welfarists should adopt. Of course, if one begins simply with the three theses above it’s hard to avoid arriving at anything other than a form of maximizing act utilitarianism. I argue, however, that the best expression of welfarism—i.e. the most compelling normative theory consistent with the three theses above—isn’t utilitarian at all. Instead, it is a kind of *ideal carer* theory. My inspiration is not a nostalgic yearning for the benevolent spectators of early proto-utilitarian history; rather it is Elizabeth Anderson, Steven Darwall, and David Velleman’s more recent insight that there is a difference between *valuing welfare* and *valuing welfare for the sake of those to whom it accrues*.[[3]](#footnote-3) Recognizing this difference allows us to develop a normative theory (as well as a “theory of beneficence”—a theory which ranks states of affairs according to welfare value) that better captures the spirit of welfarism, while avoiding some seemingly intractable problems that beset the traditional utilitarian approach. What’s more, my theory can be applied without knowing what exactly welfare consists in, or how to quantify it *intra-* or *inter-*personally.

**Why welfarism?**

A full discussion of welfarism’s strengths and weaknesses would fill a long book; I won’t attempt that here. Nevertheless, allow me to briefly outline some of the view’s charms, especially because I’ll argue the view that I propose best captures these charms.

Why accept Sumner’s welfarist thesis that “ethics has ultimately to do with ensuring that lives go well”? To begin with, this view is already deeply embedded in our moral thinking. Stereotypically wrong acts (e.g. killing, theft, rape, deception, disloyalty, etc.) seem to share one and only one obvious and unique characteristic: they each tend to be bad for their victims. Moreover, the relative seriousness of these sins tends be proportional to how bad these acts are for others. Indeed, if someone tells you you something is your duty, one might challenge the speaker by asking—“Is it? Who would it benefit? For whom is it worth doing?” Similarly, if you are told something is forbidden, one might ask—“Really? What’s the harm in it?”[[4]](#footnote-4) But these questions function as intelligible *challenges* to the original speakers only to the extent that harms and benefits are the *only* factors relevant to moral assessment. Thus, our tendency to think these questions as challenges indicates our welfarist leanings.

Of course, one might deny welfarism and hold that harm and benefit are central to moral assessment. For example, consider a deontological position consisting of a single prohibition: *Intend no harm.* One can hold such a view without being a welfarist by denying that *one’s being harmed* is bad, or by denying that the wrongness of intending harm is in any way explained by its actual or hypothetical tendency to make things worse. Either denial is obscure. What then would justify the prohibition? Surely harming is wrong in virtue of the *bad* that typically results from it?

Thus, welfarism not only provides a plausible understanding of what our moral practice is about—it is also able to *justify* it. After all, why shouldn’t we view moral norms as mere evolutionary or cultural artifacts that bear no real relevance to what I should do all things considered? Welfarists have a powerful answer: rather than insisting that moral rules have no or are not in need of justification, they insist that the point of these requirements is to make the world better in the only way it can be: by improving our lives—for our sake. In short, moral acts do what is best for everyone.

Even when welfarism profoundly conflicts with our “intuitive” judgments, its appeal is resilient. For example, welfarism’s most straightforward manifestation—act utilitarianism—may sometimes horrifically recommend hanging an innocent, torturing a child, or pushing a fat man off a footbridge. Even here, we must concede that the utilitarian has one heck of a point. If performing these acts really would be what’s best for everyone—what’s most worth doing for our “collective sake”—why shouldn’t we? Despite the objections the utilitarian welfarist faces, answering that question remains difficult. For if morality really advised us to do something other than what’s best for everyone then, as Simon Keller asks, “would we not be *better off* with something other than morality?” (2009: 91). On one reading, the answer to Keller’s question is a trivial “yes.” But surely, if there is a correct moral theory, it is not something we’d be better off without. If that’s right, welfarist theories emerge as our most natural candidates for moral truth.

**Welfarism and maximization**

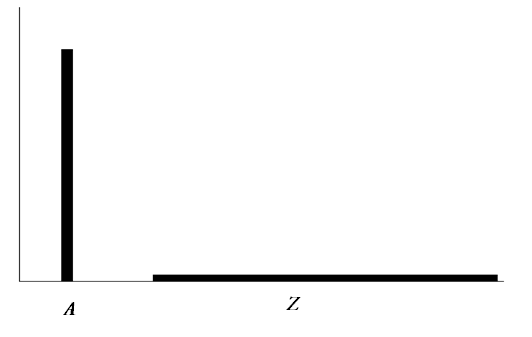
Welfarism tells us that moral norms are a function of the good; but it does not precisely identify how to “derive” these norms (Sumner 1996: 198). Welfarists hold that moral norms are (roughly) norms to “promote” well-being; but promote it when, for whom, how, and to what extent? The position, by itself, offers us no direct advice. Nevertheless, initially only one approach seems available—to *maximize* the well-being of *all* welfare subjects. Welfarism’s priority of the good tells us that facts about which states are intrinsically good explain which duties and moral reasons we have. It is difficult to see how any moral considerations that might be raised to defend, for example, a *satisficing* or *rule* utilitarianism, or a restriction to only *human* well-being could be justified by or founded in an appeal merely to the value of well-being itself. By taking the priority of the good seriously, the welfarist seems committed to no less than *maximizing* welfare.

**Maximizing welfare and population ethics**

The idea that welfare is to be maximized needs specification. It is most naturally interpreted as an injunction to maximize *total* well-being—the sum of individuals’ well-being in a state of affairs. But this view generates a repugnant conclusion that Derek Parfit famously called “the Repugnant Conclusion”:

***RC***: There is always some number *n* such that a world containing *n* creatures whose lives are barely worth living is better than any world with a smaller number of creatures who are very well-off. (1984: 381–90)

To illustrate this conclusion’s repugnance, consider the famous situations below:



Here, height represents levels of individual well-being and width the size of the population. Thus, in *A* suppose that 10 billion individuals fare 20 times better than the individuals in *Z*, but *Z* contains a population of 210 billion people. We may face a choice between situations like *A* and *Z* when we can spur or curb population growth in contexts of limited resources.

Maximization, therefore, can sometimes be an affront to the very ideas that make welfarism attractive. Choosing to move from a world like *A* to a world like *Z* certainly does not seem to be doing what’s “best for everyone”—what’s most worth doing our sake.[[5]](#footnote-5) Rather, moving to *Z* from *A* seems to make things go much worse for people. Here, unlike pushing the fat man or hanging the innocent, the characteristic pull of the utilitarian position is lost; indeed one is hard-pressed to see *any reason at all* to choose *Z*.

But our reactions to the Repugnant Conclusion do not reveal our implicit rejection of welfarism. Our view that *A* is better than *Z* is not explained by implicit deontological intuitions that challenge welfarism’s priority of the good. For were the move from *A* to *Z* to occur via some natural disaster—a procreative explosion—things still have become worse. RC also does not indict welfarism’s axiological monism. We don’t object to a move from *A* to *Z* because doing so sacrifices important values such as freedom, preference-satisfaction, or autonomy; were the beings in *A* and *Z* squirrels, and thereby incapable of manifesting these other values, the move would make things worse, and worse *for* squirrels. Increasing the number of squirrels, at great expense to their average well-being, isn’t worth doing *for their sake.* RC does not indict welfarism, it indicts expressing it in terms of maximizing *total* well-being.

But expressing welfarism in terms of maximizing *average* well-being fares no better. Simply “removing” people who are less well-off than average raises average well-being, but it does not make the world better, or better for us. To sharpen the point, imagine a world occupied by three children suffering in the 6th circle of hell. If we give birth to another occupant of this hellish world, one who suffers the mere tortures of the 5th circle, this improves average well-being, but the world is worse. Giving birth to this less tortured child wouldn’t be what’s worth doing for our or anyone’s sake—it’s not what’s “best for us.” Characterizing welfarism as an injunction to maximize average well-being betrays what makes the view compelling.[[6]](#footnote-6)

This is puzzling. Even non-welfarists tend to believe we have a moral reason to promote well-being; and welfarists believe that this is simply what morality is “all about.” But sometimes there is no reason to maximize well-being, and welfarist reasons *to avoid* doing so. And a “hybrid” theory that somehow tries to appropriately weigh both average and total well-being cannot solve the puzzle.[[7]](#footnote-7) Such views predict that we should produce “burdensome children”—children who are so happy that they promote *both* total and average welfare but reduce the well-being of everyone who existed before their birth. Arguably, the very poor are often able to produce such children; but their doing so would not be what’s best for us, or anyone at all. Similarly, such views entail that “replacing” the current population with an ever so slightly (but currently non-existent) happier population would be an improvement. Surely this is not at all what the welfarist has in mind. As Jan Narveson once put it: we care about well-being because we value making *people happy*, not making *happy people* (1976).

**Person-affecting constraints**

Narveson’s rhetoric points to a feature shared by each of these problematic cases. In each case we have an improvement in average or aggregate well-being, but no one *for whom* the outcome was better. For example, *Z* (relative to *A*) has greater total well-being, but it’s not better for anyone in particular. And when we improve average well-being by merely removing, adding, or replacing someone, there is no person for whom it is an improvement.

To avoid these problems, we might adopt a “person-affecting” welfarism.[[8]](#footnote-8) There are various ways to characterize the person-affecting approach, but the basic view may be stated as:

*Person-Affecting Constraint*: Actions or states of affairs are better or worse only insofar as they are better or worse for particular person(s).

This constraint is *too* constraining when our actions affect the identities of future individuals—Parfit calls these “Non-Identity” cases (1984: 351–379). Parfit illustrates with an example: imagine a woman who could have a child at age 14 or at 30. Suppose that the child she would have at 30 cannot be identical to the child she would have at 14, and that if she has the child at 14 it will fare worse than the child she would have had at 30. All else being equal, we tend to think it would be better if the woman had the child at 30. However, no one would be benefited by her doing so, and no one would be harmed or made worse-off were she to have a child at 14. Thus, person-affecting views are saddled with the result that it is no better if she has the child at 30.[[9]](#footnote-9) Similarly, suppose that our decision to prevent global warming also affects *who* lives in 150 years—the future population contains completely different sets of individuals depending on our decision. The person-affecting approach would seem to entail that it is no better to prevent global warming: those come to exist cannot claim to have been harmed, for they would not exist were we to have done otherwise; and if we *do* prevent global warming, those who come to exist cannot claim to have been benefited or “spared” because there’s no alternative in which they fared worse.

**An irresolvable set of problems?**

These puzzles, first noticed by Sidgwick, came into sharp focus in Derek Parfit’s *Reasons and Persons*.[[10]](#footnote-10) The problem, of course, is not whether we can find a theory of how to rank states of affairs with respect to welfare value that allows us to avoid the Repugnant Conclusion; the problem is to find a theory that can avoid the Repugnant Conclusion, the Non-Identity Problem, and any other deeply unsavory commitment. Parfit suspected that no such theory—a “Theory X”—exists, and some claim to have formally proven that such a theory is impossible.[[11]](#footnote-11) Accordingly how to deal with the Repugnant Conclusion has turned into “one of the cardinal challenges of modern ethics” (Ryberg et al. 2008).

In the years since Parfit posed these problems, various solutions have been advanced, but the consensus is that none of these proposals has been satisfying. Though I cannot give proper hearing to all of the extant theories of beneficence here, most believe that each putative solution either falls victim to a variation on one of the standard problems above, or has an unacceptable implication of its own. Many approaches have been employed and some evoke other values in addition to welfare or deontic constraints. These approaches actually give up on welfarism—the first violates axiological monism and the second violates the priority of the good. In any case, these approaches are phenomenologically inaccurate: if our verdicts in these cases were explained by goods or requirements distinct from well-being, we should at least feel *some* conflict where well-being provides us with reasons that are “dominated” or outweighed by competing evaluative or deontic concerns. Yet I can see no reason to have burdensome children, add the person to the 5th circle of hell, or prefer *Z.* Consequently, the trouble appears not to lie with other values or deontic constraints, but with something peculiar about the value of well-being itself.

If well-being was always good, we would always have some defeasible reason to promote it—either in total, average, or perhaps via some more complex process of aggregation—but we don’t. So while we may retain the welfarist idea that well-being is intrinsically good and the *only* intrinsic good, we also seem forced deny well-being’s unconditional value—we sometimes lack agent-neutral reasons to promote it. We need a principled account of *when* it matters, and hopefully one that still does justice to the spirit of welfarism. Furthermore, we don’t just want a theory that is consistent with our intuitions about cases, we want one that can explain and vindicate our intuitions. “Theory X”—if it exists—should not merely faithfully “match” our intuitions, it should capture what grounds them.

**Toward a solution**

I submit that if we want to know *when* welfare matters, we should step back and ask ourselves *why* it matters. Some theorists cannot adopt such a strategy; for them we cannot explain *why* well-being is good; *it just is*. This view has no hope of addressing the axiological problems above; but, luckily, welfarists are not condemned to it. Even if welfare is the only non-instrumental and intrinsic good, it does not follow that we can say nothing about why it has value—why it, rather than pain and suffering, is good. Welfarism only limits the *sort* of explanation we can give: *axiological monism* entails that we cannot explain welfare’s value in terms of any other *good*,and *the priority of the good* prohibits the use of deontic claims in our explanation. So is there a plausible explanation of why welfare is good? Sumner offers a suggestion:

In the case of human beings the idea that we should promote their good surely owes at least some of its appeal to the fact that we can do so for their sake . . . this notion of furthering someone’s good for her own sake can be generalized beyond the boundaries of our species; I have a perfectly good sense of what it means to take my cat to the veterinarian for her sake. (1996: 211)

On the one hand, Sumner notes something seemingly obvious and important: we value well-being because it is worth promoting *for the sake* of individuals. And indeed this may be a plausible constraint on any putative good—i.e. nothing could be good or worth promoting unless it was worth promoting for the sake of some individual(s). On Sumner’s view, welfare matters because it matters for the individuals to whom it accrues (1996: 214). If this is right, we now know we shouldn’t express welfarism in terms of simple maximization—because, as we have seen, increasing average and total welfare is sometimes not worth doing for anyone’s sake. On the other hand, we might worry that Sumner’s explanation is empty. After all, isn’t promoting an individual’s well-being just *what it is* to act for that individual’s sake? If so, Sumner’s proposed “explanation” is no explanation at all—he has expressed a tautology.

I believe Sumner is on to something: we’ve already identified a number of cases where maximizing well-being does not appear to be what is worth doing for anyone’s sake; so apparently there *is* an extensional difference between what *maximizes our well-being* and *what’s best for us*—what is most worth realizing for our sake.But this appearance cannot be vindicated, nor help us find a better theory, without identifying a corresponding conceptual difference. To appreciate the import of Sumner’s insight, and how it may lead us to an improved version of welfarism, it will help to first consider the difference between *acting for someone’s sake* and *promoting her welfare*.

**Acting for an individual’s sake vs. promoting her welfare**

What is it to act for someone’s sake?[[12]](#footnote-12) The natural answer is that acting for *x*’s sake is acting to promote *x*’s well-being. But this isn’t quite right. After all, a young child may share with or console a playmate *for that playmate’s sake* without having yet acquired a conception or theory of well-being. Indeed, care apparently precedes our particular theories of well-being and not vice versa.[[13]](#footnote-13) Furthermore, some—namely G. E. Moore, Thomas Hurka, and Don Regan—question the very intelligibility of a “good for a person,” yet we have every reason to suspect that they sometimes act(ed) for the sake of others (Moore 1903; Hurka 1987; Regan 2004).

One can also act so as to promote an individual’s well-being, but not act for her sake. Selfish foster parents who aim to promote the well-being of their foster child only so as to continue receiving government checks surely do not act for the child’s sake. Acting for one’s sake also is not acting with the *non-instrumental* aim to promote one’s well-being. Imagine a civil servant at the Department of Human Welfare who is so consumed by his work that he becomes single-mindedly concerned with, and non-instrumentally aims to promote, human well-being. Further imagine that he does not care for anybody (not even himself). Our civil servant’s obsession with welfare seems fetishistic; it is just this sort of person who might promote well-being by adding happy people or “subtracting” those with less than average well-being. And, again, these policies do not appear to be worth adopting for anybody’s sake. The difference seems to be that when we promote an individual’s welfare for that individual’s *sake*, an attitude towards that individual underlies our aim. Typically, this attitude is care. Caring for someone is incompatible with denying that her welfare matters. But our civil servant illustrates that caring for someone isn’t *merely* valuing her welfare. As appearances suggest, *care for* an individuals is an attitude we bear towards individuals, not outcomes. We don’t literally care *for* outcomes, though we can care *about* them. Nevertheless, care indirectly involves wanting certain outcomes. As Steven Darwall puts it:

Any desire for another’s good that springs from concern for that person is also for *his sake*, the object of care is the individual . . . Mill claims that people come to desire wealth even when it lacks instrumental value because of its psychological associations with other things they intrinsically desire. Were a desire for someone’s good to arise similarly, it might involve no concern whatsoever for the person himself. One might simply desire intrinsically *that* another’s good be realized without desiring it for his *sake*. (2002: 1–2)

We’re now in position to see that Sumner’s explanation, though obvious, is not vacuous. It *does* make sense to claim that we value an individual’s well-being for that person’s sake. This is just to say that we value it out of care for that individual—which is not simply caring *about* their well-being.[[14]](#footnote-14) And this explanation seems correct. We all know that people who hate themselves tend to be imprudent, and that those who hate others have little regard for anyone else’s well-being.[[15]](#footnote-15) But these cannot be conceptual truths; they’re psychological corollaries of how and why we value well-being.

This psychological explanation tells us only something about why we value well-being, not what makes it valuable. Having certain attitudes towards a person—in this case *care*—does not explain why that person’s well-being has value. If we ceased to care, their well-being would still matter. Furthermore, the explanation has a problem: we often value the well-being of distant individuals whom we may never meet—but it’s a stretch to claim we care about those *particular* people. I submit that we value the well-being of people that are unknown to us not because we care about them in particular, but because we think they are of a *type of being* that warrants or merits our care. We can care for types of individuals without caring for particular individuals of that type. And we will not view their welfare as *worth promoting* unless we view such individuals as *worth caring for*. Thus, we arrive at a corresponding metaphysical explanation for welfare’s value: *Our welfare matters only when and to the extent that we merit care.* This seems correct. Even if we hate someone, but reluctantly acknowledge that they too merit care, we’re committed to seeing their well-being as worth promoting. And we would be utterly perplexed by a utilitarian who denied that welfare subjects are worth caring for.

Elizabeth Anderson defends this view, and it later appears in the works of Stephen Darwall and David Velleman. On their shared view, personal well-being does not matter, and is not an appropriate object of moral concern, unless persons merit concern. As Anderson puts it in *Value in Ethics and Economics*:

if it does not make sense to value the person (in a particular way) then it does not make sense to care about promoting her welfare [. . .] enemies, who hate each other have no reason to promote each other’s welfare [. . .] what gives the pursuit of or desire for welfare its only point is that we ought to care about the people who enjoy it. (1993: 22)

The same position appears in David Velleman’s “A Right of Self-Termination?”

things that were good for you would not actually merit concern unless you merited concern; and if you didn’t, then despite their being good for you, they wouldn’t ultimately be worth wanting, after all. [. . .] A person’s good only has hypothetical or conditional value, which depends on the value of the person himself. (1999: 611)[[16]](#footnote-16)

And in Darwall’s *Welfare and Rational Care*:

What gives considerations of someone’s welfare or personal good the status of [agent-neutral] normative reasons is having a value that makes him worthy of care, as one accepts when ones cares for him. (2002: 8)

Given these authors’ employment history, let us call their shared position “The Michigan Condition” and characterize it as follows:

*Michigan Condition* (*MC*): Being good for some individual, *x*, is not itself a good-making feature of any state of affairs unless *x* merits care.

Of course, this principle cannot help us develop an improved welfarist theory if it is itself inconsistent with welfarism. And there may be grounds for concern. For example: can well-being have *intrinsic value* if its value depends on the value of something else—the “value” of the individuals to whom it accrues—i.e. their *meriting care*? The answer rests on two issues: what we mean by “intrinsic value” and whether facts of the form *x merits A* are evaluative facts—do they attribute value to an object? Only one combination of these views makes *MC* inconsistent with welfarism. Specifically, if (1) “intrinsic value” simply refers to value that an object has that is independent of the value of anything else, and (2) claims of the form *x merits A* are claims about *x*’s value, then *MC* *is* inconsistent with welfarism. But, as I will explain, very few accept this pair, and given the way welfarists tend to use “intrinsic value,” *MC* and welfarism are not in conflict.[[17]](#footnote-17)

A predominant and traditional use of “intrinsic value” refers to value an object has in virtue of its intrinsic properties.[[18]](#footnote-18) On this view, well-being can be intrinsically valuable even when its value depends on the value of individuals. To illustrate, if *Christian’s being well-off* is good, and *MC* is true, then it follows that my being well-off is intrinsically valuable. For were it not for my “value”—my meriting care—this state would not be good, but the state is nevertheless good in virtue of an intrinsic constituent—me. Being intrinsically good, and having a goodness that does not depend on the value of anything else, come apart at least in cases where the good state is good because of the value of an intrinsic constituent. In more rough, but vivid, terms: if my well-being *matters* because *I matter*, my faring well would still be good, as Moore would say, “in isolation.” So, given this traditional and common understanding of “intrinsic value,” *MC* and welfarism are consistent even if the value of well-being depends on the value of individuals.

But let’s sidestep tedious debates about how to best characterize a term of art, and suppose the worst-case scenario: suppose that, for whatever reason, *any* plausible view of intrinsic value must reject the possibility that an item’s intrinsic value could depend on the value of something else. Even this supposition won’t impugn the compatibility of *MC* and welfarism unless our *meriting care* entails that we *have value*. But only fitting-attitude theories of value license this entailment. Extant welfarists, as broadly consequentialist, would not accept such an account. The consequentialist injunction to *promote value*, which runs through welfarism’s various forms, becomes silly on a fitting-attitudes account. It becomes an injunction to promote items that merit particular attitudes. So if people were valuable in virtue of meriting care, we’d have some reason to produce more people even if we knew that they’d fare poorly! Clearly, a consequentialist would not believe any such thing even if they became convinced that some individuals merited care.

Thus, consequentialists apparently use “value” to refer to a different concept than those who accept fitting-attitude accounts.[[19]](#footnote-19) This result is not surprising. When the welfarist says “only well-being has intrinsic value” we may paraphrase him as saying “only states of well-being are non-instrumentally *good* or *desirable*” or that “such states are worth wanting or preferring, even if nothing else is.” Sumner, for example, characterizes intrinsic value as being “worth having or pursuing for its own sake, not merely by virtue of some further good with which it is connected or associated” (1996: 190). This is another common account of intrinsic value. And we know that it is incompatible with the view that *animals merit care* entails that *animals have value*—meriting care does not make one a suitable object of possession pursuit! Ultimately then, *MC* and the idea that well-being is the only intrinsic value are not in conflict. More importantly, given the way welfarists tend to use “intrinsic value,” such a conflict is impossible.

Nevertheless, these observations suggest a different conflict: perhaps *MC* conflicts with welfarism’s *priority of the good*. After all, if some individual(s) *merit* an attitude, isn’t it true that we *have reason* or *ought to* bear this attitude towards them? And if this is so, then shouldn’t we conclude that facts about the good are not the basis of all of our moral duties and moral reasons? This concern rests on a dubious, yet common, substantive assumption about the normative import of the *merits* relation—I call it “the deontic assumption” (or “*DA*”).[[20]](#footnote-20)

*DA*: If some attitude is “merited” or “fitting” (etc.), then we have a reason (or ought to) have this attitude.

I reject *DA*, and I think others should too. For one thing, we can all acknowledge—even consequentialists can acknowledge—that some animals are *fearsome*—that they merit or are fitting objects of fear; but then insisting that one really has a reason or ought feel fear towards such animals seems silly. After all, “what’s the point” in feeling a negative emotion provided that we can keep ourselves out of danger without it? So just as it is an open question whether you should wear a shoe simply because it fits, one can sensibly ask whether one has any reason at all to *feel* as is fit. Accordingly, *DA* is substantive normative thesis that is both in need of defense and something a defender of *MC* can reject.

More importantly, we have strong grounds for rejecting *DA*—for itmay entail that the *priority of the good* is impossible! To see this, consider two widely shared, putative conceptual truths: *being good just is* *being desirable*, and *being desirable just is* *being a* *merited/worthy object of desire*. If *DA* were true, then these apparent truths would entail that we have reasons, or ought, to desire the good regardless of whether doing so would promotethe good. This is precisely what the priority of the good denies. So *DA* allows us to dismiss traditional consequentialism and its priority of the good out of hand. Surely these positions cannot be so easily dispatched by a dubious assumption about the deontic import of the *fits* or *merits* relation. Therefore, we should doubt *DA*, and this is especially true of those who, like welfarists, accept the priority of the good. Without *DA*, the most formidable objection to pairing the Michigan Condition and welfarism goes away.[[21]](#footnote-21) [[22]](#footnote-22)

**Toward an improved welfarist theory of the good**

Our observations so far seem to be pointing in a common direction: If we accept . . .

* Welfarism is appealing because it directs us to do what would be best for everyone—to do what is most worth realizing for our sake.
* Sumner’s claim that promoting welfare is good because it is worth promoting for the sake of individuals.
* Anderson, Darwall and Velleman’s point(s) that promoting welfare for the sake of some particular person is to promote it out of care for them. And that one’s well-being would not be worth promoting unless one merited care.

. . . then it seems the welfarist theory of what states are best—what is worth wanting or realizing for the sake of welfare subjects*—*should be somehow based on “what it makes sense” to want out of care for welfare subjects. Unfortunately, how to do this is not immediately clear.

First, we might take a cue from Darwall. On his view, what is good for an individual, *I*, is what it is *rational* to desire for *I* insofar as one cares about *I* (2002: 4). Thus, our corresponding theory about what states are impersonally best—which states are most worth realizing for the sake of welfare subjects in general—would look like this:

*D*: States are good (better) to the extent that it is rational to want (prefer) them out of care for welfare subjects.

My worry about *D* is that *care*,as a kind of feeling, cannot yield commitments. Having this feeling can certainly explain why a person does something, just as feeling dizzy might, but it is at least contentious whether, like a belief, such a feeling can *rationally commit* one to anything. And insofar as it can, its commitments seem problematically agent-relative because what it is rational to do in light of an attitude is characteristically dependent on an agent’s other attitudes. For example, suppose that care for *x* commits one to valuing what one takes to be good for *x*; and suppose that I believe that a drug will improve John’s health, and you believe that the drug is bad for John’s health. Provided we both care for John, it is rational for me to want John to take the drug, while rational for you to want him to not take it. Consequently, on its face, Darwall’s view entails that what’s best for everyone would become implausibly indeterminate and “carer-relative.” Though Darwall does not intend this result, it not easy to see how to avoid it.

Darwall sometimes alternatively states his view as the view that what’s good for a person is what one *ought to want* out of care for her. Accordingly, we might characterize the welfarist impersonal good in terms of what we ought to want out of care for persons generally. But this alternative now conflicts with welfarism’s priority of the good because it makes facts about the good dependent on facts about what one *ought* to do out of care. Furthermore, this view may now be conflict with the Michigan Condition itself. Remember, Darwall claims that well-being provides no agent-neutral grounds for acting unless the relevant welfare subject(s) *merits care*. But if this is right, the natural conclusion is that there is nothing we (agent-neutrally) ought to do *merely* out of care for an individual, because the relevant individual(s) must also merit care. Thus, this alternative characterization conjoined with *MC* yields an “error-theory” about well-being. Of course, Darwall surely does not intend this entailment. Darwall may be able to explain why these concerns are misguided; my aim is not to challenge the coherence of his view. I’m only suggesting that no clear and obvious interpretation of the view can be taken “off the shelf” to assist in our search for an improved theory of beneficence.

A second suggestion is inspired by Elizabeth Anderson who writes that, “the mediating function of concern for people can never be made transparent: the appropriateness of desires for states of affairs must be subject to the constraint that they adequately express their correlative attitudes toward people” (1993: 29). Anderson, thus, appears to endorse the following principle:

A: States of individual good are valuable to the extent that realizing them would adequately express care for those individuals that merit care.

This formulation also faces problems. First, it is obscure how to determine whether an act *adequately expresses* care for B. Furthermore, because *expresses* appears to be a communicative concept, and “what communicates what” seems relative to both speakers and listeners, this view appears subject to the same sort of relativity that moved us away from the first suggestion inspired by Darwall.

Anderson’s proposal also seems to be in conflict with welfarism. It is an instance of a more general view that Anderson often invokes in *Value in Ethics and Economics*: If it makes sense to value a state, *S*, only because an individual, *x*, merits an attitude, *A*, then one should not pursue *S* through means that do not express *A* for *x* (1993: 17–43). This “expressive principle” lies behind Anderson’s ultimate view that there are deontic constraints on our promotion of our good. On her view, though it is possible to manipulatively or ruthlessly promote someone’s well-being for her sake (i.e. out of care for her), these acts don’t *express* care—indeed these types of acts communicate just the opposite. Her expressive principle does not merely tell us that we sometimes have no reason to promote well-being, it positively entails substantive norms not to promote well-being in particular ways. The resulting view denies the priority of the good, a conclusion Anderson embraces.

So Anderson’s expressive principle is questionable, and it looks like it’s in tension with welfarism. In any case, it’s certainly not obvious, and it is apparently false. I may value the state *Christian’s being on the opposite side of the street from Fido*,because Fido is fearsome (i.e. Fido merits fear), but I clearly have no reason whatsoever to only take measures to cross the street that express my fear! On the contrary, that may be what will set him off![[23]](#footnote-23)

Perhaps we’ll be more successful if we turn our attention to what care moves us to want or do, and forget about what it *makes rational*, *commits us to*, or what *adequately* *expresses* it. But here again, indeterminacy creeps in: What *I* mightprefer in light of caring for someone may be different than what *you* might, especially if you and I have different theories of well-being. Furthermore, it’s not clear the resulting theory will be suitably normative—what we *actually* prefer will be influenced by false background beliefs, poor inferences, or other idiosyncratic elements of our psychologies. These should play no part in determining what’s actually best for us. Our best option here is to idealize. Building our theory on our actual dispositions only tells us how care manifests itself in the ignorant and irrational. To avoid these distorting influences, we may invoke an Ideal Carer (“IC”)—a fully-informed, formally, instrumentally, and theoretically rational being—and ask what it would do and want insofar as it cares for welfare subjects.[[24]](#footnote-24) [[25]](#footnote-25) Of course, this option famously has indeterminacy problems of its own. Couldn’t more than one hypothetical being qualify as an IC, and yet each wants and does different things? For example, full information and formal rationality will not guarantee a convergence in views about what’s good for people—e.g. couldn’t one IC be a hedonist about well-being while the other is not?

My response is simple—the IC should have no view about what’s good for people. Indeed, our IC should have *no attribute* that is not necessitated by our description of it. Any addition would either be (a) superfluous, if it makes no dispositional difference, or, (b) yield “junk” outputs if it does make a dispositional difference. If we want to identify the purest, least-arbitrary, and immaculate manifestations of care, then we should not add anything that might make a dispositional difference that is not directly attributable to its care or the idealizing conditions themselves. So unlike Firth’s “Ideal Observer” (1952), who does not have the relevant care, and is “otherwise normal,” the IC has no desire to sing while listening to the radio—normal though it may be. So, the proper view of the IC is not as *any* or *a* being that cares for all welfare subjects and is fully-informed (etc.). Instead, the IC is the *least* *complex* being who meets these conditions. With that clarification, the problem of “many” ICs with different dispositions goes away.

In short, using the IC gives us the most plausible way to model what care “recommends.” Care, obviously, does not *literally* recommend anything, and similarly it’s a stretch (and perhaps even a category error) to say it renders some actions intelligible or reasonable—that there are acts that *make sense* or that *one ought to* *perform* in light of caring. And few acts beyond hugs, facial expressions, and speech-acts actually express care, if this means anything beyond providing evidence for its presence. But we can, and should, turn our attention to the *non-arbitrary* and *error-free* outputs or manifestations of care. The IC’s idealization aims to ensure these outputs are not infected by error or ignorance. Its minimal complexity ensures the outputs are not arbitrary.

Of course, now one might worry that the IC cannot translate its care into action—a caring person tries to do what’s best for those for whom they care, but with no theory of well-being, the IC’s care cannot recommend any response. This objection is empirically implausible. We care for people, but almost none of us have a clear or robust theory of what is best for the people for whom we act. Again, children and G. E. Moore can act from care, even if they haven’t the faintest sense of what the good for us is supposed to be. Yes, caring people want what is best for those for whom they care. But if that is so, I cannot see how the caring person, when formally rational, fully and vividly informed about the non-normative, and devoid of all normative views (for example a theory of well-being!) will fail to hit its mark—what they desire actually *will* be what’s good for the person. What further information could the IC require? The IC’s dispositions would at least appear to be a perfectly reliable *indicator* of what is best for us; this is does not commit us to Darwall’s stronger claim that welfare *just is* what it makes sense to want insofar as you care for someone.[[26]](#footnote-26)

The alternative—that the IC’s responses *would not* track what is best for us—seems mistaken. It is not merely that children, some non-human mammals, and G. E. Moore are capable of acting from care without a theory of well-being, it is also true that when they act from care they can do a reasonably good job of wanting what actually is good for those for whom they care. In fact, we are inclined to attribute care *because of* successful sensitivity to those states that promote an individual’s well-being. For example, we are tempted to infer that mother bears have feelings of care for their cubs because they tend to do things that promote their cubs’ well-being. But the bear, we assume, lacks judgments about what well-being is, and how to measure it. Consequently, we assume that care (plus adequate non-normative information) is sufficient to allow you to be sensitive to what is actually good for an individual.

Think of it this way, imagine that you have a difficult choice to make, and you wonder which would be *best for* you. You consult two people. First, you consult a good friend. She is intelligent and cares deeply for you. She knows you and your circumstances rather well. However, your friend has not thought much about well-being before. Indeed, were she to study the matter she might, like many of us, find herself unable to settle on a view—shifting from varieties of hedonism, to informed-desire views, to various objective-list views and back again. Alternatively, you can consult your colleague—a leading scholar with a considered substantive view about well-being. Your colleague knows you and your circumstances quite well, but does not care for you. He is willing and able to faithfully apply his theory to your decision. Who would you consult? Whose advice is likely to be more reliable? The IC, like your friend, does not need a theory of well-being, or views about how to measure it.

This points to an important advantage in using the IC to characterize what is best for us—it requires no substantive assumptions about what is good for us, how to measure it, or how to do interpersonal utility comparisons. Indeed, I submit that determining the IC’s dispositions would finally allow us to settle those issues; they would reveal what well-being is, and what is best for welfare subjects in general. But our welfarist axiology must do more than identify “good” states—it also needs to *rank* them. Accordingly, we should look to the IC’s *preferences*:

C: State, S, is better than a state, S’, iff S would be preferred to S’ by a minimally complex IC insofar as he cares for welfare subjects.

**Will this help with the problems in population ethics?**

I’ve offered an abstract case for using the IC’s preferences to refine our welfarist axiology. But the proof is in the pudding—will this theory really avoid the cases that cast doubt on the maximizing approaches?

First, notice that caring persons will not prefer policies that improve total or average well-being merely by adding people. Caring for a type of creature does not involve preferring more of them. For example, most of us care, and think it is appropriate to care, for children. Obviously, if you care for children you’ll want the children who come to exist to be as happy as can be, and you will disvalue outcomes where children suffer. But care for children does not involve preferring *more* of them, even if we’re certain that the additional children will be happy. Or consider our care for fictional characters: caring for Huckleberry does not incline me to want a happy version of him, or someone like him, to actually exist. Therefore, this view, like the otherwise problematic “person-affecting” views, can capture and explain why we ought to make people happy, not happy people. The proposal can also similarly explain why we have no standing reason to reproduce, or to create burdensome children even if the addition of such children would raise total or average well-being.

The view also avoids the problems that beset the ‘person affecting’ approach. But to see why we need to first discuss *how* our IC cares for welfare subjects. Notice that we can either (1) care only for the particular individuals that happen to meet a description, or, (2) we can care for individuals (whatever their particular identities, if any) under a description. Anderson holds that properly expressed care for people is directed only towards *each* *particular* person, for otherwise it simply amounts to a perverse concern for well-being itself, and the Repugnant Conclusion looms (Anderson 1993: 27–29, 35). But these are not our only options. To illustrate, an amnesiac woman—upon discovering scars from an apparent C-section—might be struck by an apparent and profound feeling of care for *her children*, even if it turns out that she does not have any. So does she care for her children or not? In the first sense she does not, and in the second sense she does. We might say that the woman has a *de dicto* care for her children rather than a *de re* attitude directed at particular individuals that meet that description. But this attitude towards *her children* gives her no reason to *have* a child, even if she knows it will fare well.

Notice also that individuals do not merit care because of their particular identities—e.g. because one is Bob, Suzy, or Gary. *Because Bob is Bob* is not a reason to care for him. Rather, we think that persons merit care not because of *who* they are, but *what* they are. Thinking otherwise violates the putative conceptual truth that evaluative facts cannot vary without a relevant qualitative difference. For the pure welfarist, the relevant feature will not be *being a particular person at t*, or even *being human*, but rather *being a welfare subject*—being an individual for whom things can go better or worse. Our characterization of the IC should reflect this fact; doing so allows us to handle the non-identity cases. The IC does not simply care for all the particular individuals who happen to be welfare subjects (e.g. Bob, Suzy, and Gary); rather the IC has *de dicto* care for *welfare subjects*, just as the amnesiac cares for the children she may or may not have.[[27]](#footnote-27)

Now consider the older/younger mother case that introduced the non-identity problem. On its face, we now have every reason to suspect that the IC will prefer the outcome where the mother has the child at 30 to the outcome where she has it at 14*.* As argued above, caring agents don’t prefer the addition of happy individuals, but they do want whoever comes to exist to be as well-off as possible. Apparently, an agent who cares for welfare subjects and not merely for the particular individuals that happen to be welfare subjects, would prefer Child at 30 to Child at 14*.* Were someone indifferent between these options, it would be reasonable to ask “don’t you care about your children?” Thus, we need not appeal to the otherwise unacceptable claim that welfare is independently good to explain the non-identity cases. And though it initially sounds odd, we can now see how a state may be worth realizing for welfare subjects without being worth realizing for the sake of any *particular* welfare subjects.[[28]](#footnote-28) Importantly, though this move “gets us out of a jam” it’s not *ad hoc*; rather it is the most sensible move when we consider *why* individuals possess moral standing, and why we value well-being. Earlier approaches tried to avoid deeply counter-intuitive implications with specific suggestions as to *whose* welfare matters (actual individuals, future individuals, possible individuals . . . etc.) and *how much*—each answer beset with its own set of problems. Appealing to care allows us to avoid this. Like the amnesiac mother, it is possible to qualify as caring for individuals under a description without caring for any particular entities—actual or merely possible. Caring does not require *there being* particular individuals that are the object of care. Some children care for unicorns or imaginary friends, though such beings may be metaphysically impossible.

So the theory looks promising: thus far, the dispositional profile for care corresponds perfectly to our verdicts about when welfare is worth promoting. Nevertheless, these observations only point to a research program; and some may worry that any such program is doomed. After all, Parfit suggests that, and others claim to have formally *proven*, that a satisfactory axiology for welfare is impossible (Parfit 1984: 419) (Arrhenius 2001, 2011). The proofs begin by listing extremely plausible constraints on any defensible theory. In Parfit’s informal proof, one constraint is that the theory must not entail the Repugnant Conclusion—e.g. it must not entail that *Z* is better than *A*. The proofs then purport to show that when we compare states of future generations with different sized populations, the conjunction of these constraints is impossible. The upshot is that our reasoning about value appears to be incoherent, and that perhaps we should conclude that there are no *facts* about what makes the world go better or worse for us (Arrhenius 2011: 19–23).

If these proofs succeed, my theory must violate one of the adequacy conditions, thereby “proving” it is intuitively unacceptable, incoherent, or both. In fact, it’s actually *very* easy to show that my view violates some of these putative constraints. To see this, consider *A* and *Z* again. In an abstract sense, if we were like a caring deity, and had to choose between *A* or *Z*,it certainly seems that all else being equal, we would prefer *A.* This seems like good news, because it suggests the theory will avoid the Repugnant Conclusion. But not so fast. The IC’s preference will change when it is not (like a deity) choosing between *A* and *Z* *ex nihilo*. For example, suppose that the prevailing state of affairs were *Z*. Would things *go better* if we stayed in *Z* or immediately changed our world into *A*? Would a caring person prefer that almost everyone be annihilated, so that a (relatively) small number could be much happier? No. If the prevailing state of affairs were *Z*, then surely the carer *would* prefer *Z* to *A*. But if *A* were the prevailing state of affairs, then a carer would prefer *A*—caring people don’t prefer the addition of many poorly-off people, especially when adding them would make everyone else much worse off. Thus, my theory yields the following results:

*A* is better than *Z*, if the current world is *A.*

*Z* is better than *A*,if the current world is *Z.*

The lesson we are *supposed* to draw from these results is that the theory is repugnant and incoherent. After all, the theory predicts both the repugnant verdict that *Z* is better than *A*, and *even worse*, it then later predicts that *A* is better than *Z*! The proofs typically assume no theory could have such “unacceptable” outputs. But, I hope you’ll agree that there’s nothing wrong with these results; these are the right results, something is wrong with the constraints.

The primary concern about this view would be that it makes the *better than* relation intransitive. But that is not so. Yes, a carer’s preference for *A* or *Z* is sensitive to facts about the current world, and this makes it possible for the relative value of *A* and *Z* to change. Nevertheless, the variation in the carer’s particular preference orderings makes perfect sense. The orderings depend on stable and transitive preferences regarding *types of changes*. For example, when the view predicts that:

*A is better than Z*,when the current world is *A—*because the IC prefers no change (*A to A*) to a change from *A to Z*.

*Z is better than A*,when the current world is *Z*—because the IC prefers no change (*Z to Z*) to a change from *Z to A.*

*A is better than Z*, when no one currently exists—because the IC prefers the change from an empty world to *A* to the change from an empty world to *Z*.

It is *these* preferences—preferences for types of changes—that order the IC’s further preferences for worlds *at a time*. The lesson *is not* that the “better than” relation is intransitive; it is that states of the world *at a time* have value only in virtue of the value of changes that they instantiate. The more fundamental evaluative relation holds among changes, and here, the *better than* relation remains transitive. This type of view, one excluded by the terms of the current debate, offers the most faithful rendering of how we value well-being. I believe that only this type of view will properly explain why, for example, the world goes better *when our lives are improved*, rather than *replaced with improved lives*—for these are claims about the value of different changes, not states at a time.[[29]](#footnote-29)

**Using the IC in a welfarist normative theory**

So far, we’ve used the Ideal Carer to rank and evaluate outcomes—to deliver a refined welfarist axiology*.* Effectively, the theory tells us which outcomes are best (and better) for us—which outcomes are most worth realizing for our sakes. But to get a refined welfarist *normative theory*, we still need to decide how to characterize our moral obligations in light of the good—for example, whether we should adopt an act or rule consequentialist theory. But I think we can sidestep, and potentially resolve, debates about how welfarists should characterize our obligations by again appealing to the IC. The IC’s various types of disposition will offer us a natural way to distinguish the various types of moral assessment, including the *right*, the *merely permissible*, the *morally optimal*, and the *supererogatory*. Traditional (consequentialist) approaches famously have trouble finding a principled account of these different categories. Using the IC helps us carve out the appropriate conceptual distinctions while allowing for the substantive possibility that there are no supererogatory acts. For example, the *wrong* designates a set of acts that, morally speaking, the will mustn’t perform—here unlike the *supererogatory*, one does not have the moral option to sacrifice one’s virtue. Thus, if the IC embodies the moral perspective, we may characterize the *wrong* in terms of what it would *will* that you *not do* (i.e. cases where it would override your freedom to choose for yourself), the *right* in terms of what it would *will* that you *do*, and the *supererogatory* as perhaps just the class of actions that it would prefer that you do, but would not *will* that you do. The issue of whether there are any supererogatory actions is open and depends on whether the IC *always wills* that you do what it *wants* you to do; an affirmative answer would vindicate a view akin to traditional act consequentialism.

So, as a first approximation, the best expression of welfarism may look something like this:

*The Morally Optimal:* An action, *x*, is **morally** **optimal** in a circumstance *C*, for an agent *A*, iff a minimally complex, formally rational, fully-informed being would most want *A* to *perform* *x* in *C* insofar as it cares (“de dicto”) for welfare subjects.

*The Morally Required: A* has a **moral duty** to do *x* in *C* iff a minimally complex, formally rational, fully-informed being would *will* (not merely want, hope, or wish) that *A* do *x* in *C* insofar as it cares for welfare subjects.

*The Supererogatory:* X is **supererogatory** for *A*in C iff a minimally complex, formally rational, fully-informed being would most *want* but not *will* that *A* do *x* in C insofar as it cares for welfare subjects.

*The Permissible:* X is **permissible** for *A*in C iff a minimally complex, formally rational, fully-informed being would not *will* that *A* not do *x* in C insofar as it cares for welfare subjects.

One might suspect that these standards run afoul of welfarism’s priority of the good—notice that none of them make reference to the good or to welfare! But this is misleading. On closer inspection, these standards positively entail the priority of the good. After all, the IC’s dispositions will be some function of his *desires* and *preferences*—the dispositions that correspond to our axiological theory. Less abstractly, it seems that a caring person has no basis for willing that anyone do anything unless there are beings whose lives could be made better or worse—so it is well-being that provides the ground for his dispositions.

**But is this theory of any practical use?**

Whether this is “best expression of welfarism” or not, it’s interesting in that we’ve arrived at a pure and uncompromising welfarism that is not necessarily utilitarian. However, one might worry that the theory suffers from unique and intractable epistemic hurdles. Specifically, we might worry that it is useless as a guide for discerning what is right, wrong, optimal, or best for us, because we cannot access the IC’s dispositions. This concern is less serious than it seems. My proposal, I submit, substitutes a tractable and empirically investigable research question—*what is the dispositional profile of sympathetic concern?—*for the more difficult and perhaps intractable questions of *what is well-being?* And *how can it be quantified and measured within and among individuals?* Care is familiar to us; we’ve each felt it for others. And though the term ‘care’ is ambiguous, the affective state that I (and Darwall) appeal to is also an object of psychological study.[[30]](#footnote-30) Presumably, we can identify those who have care for others and those that do not. Identifying both will be useful—for we can compare the dispositional profile of those who care for others versus the profile for those who do not. Obviously, none of our subjects will be “minimally complex.” But I presume that with careful experimental design and a large and diverse sample size, we can begin to identify which responses among the caring are attributable to their care. Advances in neurology may even allow us to identify which brain-states or processes are associated with care, and which response-types it prompts.

Finally, given that none of us are fully informed, we might worry that knowing what care for welfare subjects disposes *us* tells little about the dispositions of the IC. But actually it tells us an awful lot. It tells us what the IC would do if the circumstances were as we take them to be. And we can control for differences among subjects’ non-normative background beliefs by asking them to respond to detailed thought experiments, or scenarios where the circumstances are stipulated. Control for the influence of different normative background beliefs could be achieved with a large and diverse set of subjects. Generalizations should emerge, and these may function as moral or axiological principles depending on the relevant response-type. Of course, we should be careful about pronouncing from the armchair which conclusions we might reach. I’ll leave open a deep meta-ethical question about these principles: whether or not the IC *tracks* these truths or its responses *make* them true. For practical purposes, that question is irrelevant. But the view is not an ideal-observer or response-dependent theory “all the way down” because it is built on the independent moral assumption that those whose lives can go better or worse merit care.

Of course, even if this research project is viable, there is only so much we can know about the IC’s dispositions. Nevertheless, if welfarism is true, the theory gives us a moral ideal that we can approximate by: (1) cultivating impartial concern for sentient creatures; (2) isolating and acting only upon that affective response; and (3) cultivating the epistemic virtues. The extent to which you can accomplish 1–3 is the extent to which you and your actions approximate the moral ideal. Perhaps that’s all we need: perfection is for the idealized; improvement is all we can hope for here on Earth.

REFERENCES

Anderson, E. (1993) *Value in Ethics and Economics* (Cambridge, MA: Harvard University Press).

Arrhenius, G. (2000) “Future Generations: A Challenge for Moral Theory.” FD Diss., Uppsala University, Dept. of Philosophy (Uppsala: University Printers).

—— (2011) “The Impossibility of a Satisfactory Population Ethics.” Unpublished manuscript.

Batson, D. (1991) *The Altruism Question: Toward a Social-Psychological Answer* (Hillsdale, NJ: Lawrence Erlbaum Associates).

—— (2009) “Two Forms of Perspective Taking: Imagining How Another Feels and Imagining How You Would Feel.” In K. D. Markman, W. M. P. Klein, and J. A. Suhr (eds.), *Handbook of Imagination and Mental Simulation* (New York: Psychology Press), 267–279.

—— and Shaw, L. (1991) “Evidence for Altruism: Towards a Pluralism of Pro-Social Motives,” *Psychological Inquiry* 2, 107–122.

Bradley, B. (2002) “Is Intrinsic Value Conditional?” *Philosophical Studies* 107, 23–44.

—— (2006) “Two Concepts of Intrinsic Value,” Ethical Theory and Moral Practice 9, 111–130.

Broome, J. (1999) *Ethics Out of Economics* (Cambridge University Press).

—— (2004) *Weighing Lives* (Oxford University Press).

Coons, C. and Weber, M. (2011) “Transitional Consequentialism and the Mere Addition Paradox.” Unpublished manuscript.

Darwall, S. (2002) *Welfare and Rational Care* (Princeton University Press).

Driver, J. (2009) “The History of Utilitarianism.” The Stanford Encyclopedia of Philosophy (Summer 2009 Edition), ed. Edward N. Zalta <http://plato.stanford.edu/archives/sum2009/entries/utilitarianism-history/>

Feldman, F. (1997) Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy (Cambridge University Press).

Firth, R. (1952) “Ethical Absolutism and the Ideal Observer,” *Philosophy and Phenomenological Research* 12, 317–345.

Hare, C. (2007) “Voices From Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?” *Ethics* 117, 498–523.

Heathwood, C. (2003) Welfare and Rational Care (Review), Australasian Journal of Philosophy 81, 615–617.

Heyd, D. (1988) “Procreation and Value: Can Ethics Deal With Futurity Problems?” Philosophia 18, 151–170.

Hoffman, M. (1981) “Is Altruism Part of Human Nature?” *Journal of Personality and Social Psychology* 40, 121–137.

—— (1991) “Is Empathy Altruistic?” *Psychological Inquiry* 2, 131–133.

Huemer, M. (2008) “In Defence of Repugnance,” *Mind* 117, 899–933.

Hurka, T. (1987) “‘Good’ and ‘Good For’,” *Mind* 96, 71–73.

Jaworska, A. (2007) “Caring and Internality,” *Philosophy and Phenomenological Research*, 74(3), 529–568.

Keller, S. (2009) “Welfarism,” *Philosophy Compass* 4(1), 82–95.

McMahan, J. (1998) “Wrongful Life: Paradoxes in the Morality of Causing People to Exist.” In Jules Coleman and Christopher Morris (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka* (Cambridge University Press), 208–247.

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

—— (1951) *Philosophical Studies* (Atlantic Highlands, NJ: Humanities Press).

Narveson, J. (1976) “Moral Problems of Population.” In M. D. Bayles (ed.), Ethics and Population (Cambridge, MA: Schenkman), 59–80.

Parfit, D. (1976) “On Doing the Best for Our Children.” In M. D. Bayles (ed.), Ethics and Population (Cambridge, MA: Schenkman), 100–115.

—— (1982) “Future Generations: Further Problems,” Philosophy and Public Affairs 11, 113–172.

—— (1984) Reasons and Persons (Oxford: Clarendon Press).

Parsons, J. (2002) “Axiological Actualism,” Australasian Journal of Philosophy 80(2), 137–147.

Rachels, S. (2004) “Repugnance or Intransitivity: A Repugnant but Forced Choice.” In J. Ryberg and T. Tännsjö (eds.), *The Repugnant Conclusion: Essays on Population Ethics* (Dordrecht: Kluwer), 163–186.

Rawls, J. (1971) *A Theory of Justice* (Cambridge, MA: Belknap).

Regan, D. (2004) “Why Am I My Brother’s Keeper?” In R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value: Themes from the Moral Philosophy* *of Joseph Raz* (Oxford University Press), 202–230.

Roberts, M. (1998) Child versus Childmaker: Future Persons and Present Duties in Ethics and the Law (Lanham, MD: Rowman & Littlefield).

—— (2004) “Person-Based Consequentialism and the Procreation Obligation.” In J. Ryberg and T. Tännsjö (eds.), *The Repugnant Conclusion: Essays on Population Ethics* (Dordrecht: Kluwer), 99–128.

Rosati, C. (2006) “Darwall on Wefare and Rational Care,” *Philosophical Studies* 30, 619–635.

Ryberg, J., Tännsjö, T., and Arrhenius, G. (2008) “The Repugnant Conclusion.” The Stanford Encyclopedia of Philosophy (Fall Edition), ed. Edward N. Zalta <http://plato.stanford.edu/archives/fall2008/entries/repugnant-conclusion/>

Sider, T. (1991) “Might Theory X be a Theory of Diminishing Marginal Value?” Analysis 51, 265–271.

Sidgwick, H. (1907) *The Methods of Ethics*, 7th edition (University of Chicago Press).

Sikora, R. I. (1975) “Utilitarianism: The Classical Principle and the Average Principle,” Canadian Journal of Philosophy 5, 409–419.

Sumner, L. M. (1996) *Welfare, Happiness and Ethics* (New York: Oxford University Press).

Tännsjö, T. (2002) “Why We Ought to Accept the Repugnant Conclusion,” Utilitas 14, 339–359. Reprinted in J. Ryberg and T. Tännsjö (eds.), *The Repugnant Conclusion: Essays on Population Ethics* (Dordrecht: Kluwer), 219–238.

Temkin, L. (1987) “Intransitivity and the Mere Addition Paradox,” Philosophy and Public Affairs 16, 138–187.

—— (1993) “Harmful Goods, Harmless Bads.” In R. G. Frey and C. W. Morris (eds.), Value, Welfare, and Morality (Cambridge University Press), 291–324.

Thompson, R. A. (1987) “Empathy and Emotional Understanding: The Early Development of Empathy.” In N. Eisenberg and J. Strayer (eds.), *Empathy and its Development* (Cambridge University Press), 119–145.

Velleman, J. D. (1999) “A Right of Self-Termination?” *Ethics* 109, 606–628.

—— (2008) “Beyond Price,” *Ethics* 118, 119–212.

Warren, M. A. (1978) “Do Potential People Have Moral Rights?” In R. Sikora and B. Barry (eds.), *Obligations to Future Generations* (Philadelphia: Temple University Press), 14–30.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., and Chapman, M. (1992) “Development of Concern for Others,” *Developmental Psychology* 28(1), 126–136.

1. 1 My thanks to the anonymous reviewers, Connie Rosati, Pekka Väyrynen, Kelly Sorensen, Simon Keller, David Faraci, Mark Timmons, David Schmitz, Ellie Mason, Doug Portmore, David Sobel, David Shoemaker, Chris Heathwood, Guy Fletcher, Elizabeth Guthrie, and Tom Dougherty for comments on this paper. [↑](#footnote-ref-1)
2. Keller (2009) characterizes welfarism slightly more broadly. [↑](#footnote-ref-2)
3. Perhaps the most important difference between the type of theory endorsed by Hume and Smith and the one I will defend here lies in the relevant emotional response. For Hume and Smith a kind of empathy is the relevant sentiment. But “care” is a different, but related, state that Darwall (2002) also calls “sympathetic concern.” This particular state, as distinguished from types of empathy, has become an object of psychological study: see Hoffman (1981, 1991), Batson (1991, 2009). [↑](#footnote-ref-3)
4. Some rhetoric here is cribbed from Sumner (1996: 191). [↑](#footnote-ref-4)
5. Some, for lack of better options, now embrace the RC; see: Huemer (2008), Broome (2004), and Tännsjö (2002). [↑](#footnote-ref-5)
6. This sort of example, I believe, first appears in Parfit (1984: 422). [↑](#footnote-ref-6)
7. Sidgwick (1907) may be the first to suggest such a view—see Driver (2009). [↑](#footnote-ref-7)
8. See Narveson (1976). For a more sophisticated contemporary approach, see Roberts (1998) and (2004). [↑](#footnote-ref-8)
9. A more developed critique of person-affecting views appears in McMahan (1998). [↑](#footnote-ref-9)
10. Sidgwick (1907) and Parfit (1984). [↑](#footnote-ref-10)
11. See, for example, Arrhenius (2011) and Rachels (2004). [↑](#footnote-ref-11)
12. The contribution of ‘sake’ to expressions of this form isn’t clear: “I did it for John” and “I did it for John’s sake” seem to be equivalent except in the rare case where the former expresses that one did something in John’s stead. [↑](#footnote-ref-12)
13. See Jaworska (2007) and Darwall (2002: 50–72); for evidence from experimental psychology see Thompson (1987) and Zahn-Waxler et al. (1992). [↑](#footnote-ref-13)
14. For concerns about Darwall’s claim that we can characterize care independently of well-being, see Heathwood (2003: 616–17). [↑](#footnote-ref-14)
15. For similar arguments, see Darwall (2002: 6) and Velleman (1999: 611). [↑](#footnote-ref-15)
16. Velleman’s view has changed slightly. In “Beyond Price” (2008), he departs from Anderson and Darwall by claiming that the particular emotional response that’s relevant to well-being is *love*. For more concerns about sympathetic concern as the attitude that’s appropriately relevant to a person’s good, see Rosati (2006). [↑](#footnote-ref-16)
17. Only Anderson, to my knowledge, accepts this pair. Although I don’t think this puts her in *substantive* dispute with others. Her commitment to (2) is apparent in her well-known contention that the value of states of affairs depends on the value of individuals (Anderson 1993: 17–43). Her commitment to (1) is revealed when she describes extrinsic values as items it makes sense to value only because it makes sense to value something else (Anderson 1993: 19). [↑](#footnote-ref-17)
18. See Moore (1951: 260) and Feldman (1997: 136–39). For a defense, see Bradley (2002). [↑](#footnote-ref-18)
19. For a compelling argument along these lines, see Bradley (2006). [↑](#footnote-ref-19)
20. I’ve used the expression ‘merits care’ but, to my mind, the expressions ‘worth caring for’, ‘fitting object of care’, or ‘appropriate object of care’ serve just as well, and often refer to the same concept. Although ‘merit’ sometimes connotes that the relevant individual deserves (in the sense that they have *earned*) the relevant attitude. I do not intend that sort of reading; I use ‘merits’ because it is the shortest. [↑](#footnote-ref-20)
21. One might worry that my conclusion is too weak: perhaps *MC* is consistent with welfarism, but that does not show that it is a welfarist view! Indeed, neither Velleman, Anderson, nor Darwall happen to be welfarists. But hold your horses, I don’t claim that *MC* is welfarist, I use it to develop a welfarist view (below). [↑](#footnote-ref-21)
22. Not only is it dubious whether these relations entail deontic relations to do/feel as it fit, entailments also do not go in the other direction. To illustrate, imagine a world populated by infants. Surely these infants merit care. But it is false that they *ought* to be cared about—no one in the world is cognitively equipped to care about them—so no one *can* care about them. Meriting care is also not being such that one ought to care, *if one can*. If I promise a plant-loving friend that I will really care for (and not merely tend to) his plants while he is away, then I ought to care for his plants. But that fact does not entail that the plants merit care. My own view about how fitting attitudes are normative, without being deontic or evaluative in the ordinary sense, will become clearer below. In short, on my view, the special normative role of the merited attitudes are as the conative elements that supplement *formal* ideal practical deliberation—they are what to attribute to any ideal observer used as standard or heuristic for resolving practical questions. [↑](#footnote-ref-22)
23. One might object that I’m equivocating on the meaning of ‘express.’ Perhaps one might say that, *A* can express *I* in the sense that (i) i*s motivated*, or is *guided* *by*, *I*, or in the alternative sense where *A* (ii) can *display,* or *communicate* the relevant attitude. But Anderson clearly seems to have the later in mind (see Anderson, 1993: Chapters 1–2, especially pp. 3, 11, 18). Anderson believes that if we a looked just at what one would *want* insofar as one were impartially caring, we would again arrive at the *repugnant conclusion* because impartial care would lead one to prefer *Z* for the sake of the people who do not exist in *A*: these people would be “better off” in *Z*, because non-existent in *A* (1993: 28). I think Anderson is wrong about this. First, she seems to dubiously assume that bringing into existence (with a life “worth living”) is a benefit. Second, her conjecture is phenomenologically inaccurate: as a carer I do not feel pangs of conflict in thinking about how we might create even more people. The carer just is, all else being equal, indifferent to “mere addition.” Her view, instead, appears to be the odd suggestion that care for people *moves* one to add more people, but doing so is somehow in conflict with how one meaningfully *expresses* one’s care. [↑](#footnote-ref-23)
24. More precise characterizations of the relevant idealization will need to be made. Here, I’m merely sketching, not mucking through the details. However, an “insofar as” clause is needed because the IC will have some dispositions merely in virtue of being formally rational and fully-informed. We will not want these dispositions to deliver moral verdicts—e.g. the formally rational person draws valid inferences, but it’s not plausible to thereby claim that doing so is morally optimal. The “insofar as” clause ensures the relevant response would not have occurred *but for* the IC’s care. [↑](#footnote-ref-24)
25. Darwall considers a position like this but opts instead for the “normative” view that well-being is what one *ought* or what it *makes sense* to want insofar as one cares (2002: 31). He appears to move away from a view like mine because it would fail to capture well-being’s normative character—a caring person could recognize that certain states would be preferred by an IC, but ask “So what? How does that bear on what I should prefer?” I’m not moved by this worry for two reasons. First, caring agents presumably will be committed to approximating the preferences of the IC, because they are committed, as agents, to not acting on faulty or incomplete information and to meet the formal constraints of practical and theoretical reasoning. So they really cannot coherently ask, “What’s it to me?” Second, my aim is not to provide an account of the *concepts* of well-being or even what’s best for us, but to give a substantive characterization of which states fall under the extension of these concepts across worlds. [↑](#footnote-ref-25)
26. A concern about this view is that care for individuals may involve dispositions which clearly cannot be used to characterize what is good for them. For example, caring for others may often dispose us to tell others that we care, but telling someone that you care may not be part of what’s good for them. In short, caring persons have other dispositions that go beyond wanting or preferring what is in fact good for the person. If so, an IC’s wants or preferences would not reliably indicate the caree’s good. But I submit that desires to communicate one’s care, etc. are often founded on values, desires, and aims other than pure care itself. Indeed, it may not be possible for an attitude to be partly constituted by disposition to communicate itself; for a regress looms. Consequently, while what actual caring people tend to want for others may not be a reliable indicator of person’s good, a minimally complex IC’s desires for others would remain so. [↑](#footnote-ref-26)
27. My description of care here is, in one respect, a radical departure from the way it is discussed by both Anderson and Darwall. Specifically, they often characterize it as directed at particular individuals, and obviously I deny that. Kelly Sorensen brought this to my attention. However, this does not mean we’re talking about different affective states under the same name. Rather, I think the cases above (caring for types of beings and impossible beings) illustrate that the very same affective state can have no particular beings as its object. The IC *does* care for Gary, Suzy and Bob, but it does so in virtue of its more general attitude toward beings of their type. [↑](#footnote-ref-27)
28. See Hare (2007) for a related move. Hare also appeals to the distinction between particular individuals that happen to meet a description and individuals under a description. My argument appeared prior to the publication of his article in my 2006 dissertation. In conversation, I learned that he also developed his argument in *unpublished* form before I did. [↑](#footnote-ref-28)
29. Obviously, more needs to be said about this approach, but space limits us here. For an extended defense and articulation of this approach, see Coons and Weber (2011). [↑](#footnote-ref-29)
30. See Hoffman (1981, 1991), Batson (1991, 2009), and Batson and Shaw (1991). [↑](#footnote-ref-30)