PREFERENTISM AND SELF-SACRIFICE

BY

CHRIS HEATHWOOD

Abstract: According to the argument from self-sacrifice, standard, unrestricted desire-based theories of welfare fail because they have the absurd implication that self-sacrifice is conceptually impossible. I attempt to show that, in fact, the simplest imaginable, completely unrestricted desire-based theory of well-being is perfectly compatible with the phenomenon of self-sacrifice – so long as the theory takes the right form. I go on to consider a new argument from self-sacrifice against this simple theory, which, I argue, also fails. I conclude that, contrary to popular opinion, considerations of self-sacrifice do not pose a problem for preferentist theories of welfare.

According to one of the leading accounts, human welfare has fundamentally to do with desire – with getting what one wants. This paper concerns a popular line of argument against this approach, which contends that it goes wrong when it comes to self-sacrifice. In an influential article devoted to this objection, Mark Carl Overvold writes

> . . . that if we identify an agent's self-interest with what he most wants to do, all things considered, it becomes logically impossible that there ever be a genuine instance of self-sacrifice. . . . (Overvold, 1980, p. 117)

Quite a few philosophers of welfare agree. Richard Brandt, a one-time defender of the desire approach,[1] came to accept Overvold's criticism, agreeing that

> . . . as Professor Mark Overvold points out in a recent paper, if we are to make a distinction between self-interest and self-sacrifice, I must have some desires that it is not in my self-interest to satisfy, and hence desires the satisfaction of which does not, as such, add to *my* welfare or utility. (Brandt, 1982, p. 173)

Pacific Philosophical Quarterly 92 (2011) 18–38

Crediting Overvold, James Griffin also agrees that

> . . . the desire account . . . has difficulty distinguishing between selfish and selfless action. (Griffin, 1986, p. 316, 25n)

Stephen Darwall is another philosopher of welfare who cites Overvold approvingly, writing that the view that

> . . . what is *in* a person's interest . . . consist[s] in whatever an agent (rationally) *takes* an interest in . . . make[s] self-sacrificial acts . . . impossible. (Darwall, 2002, p. 24)

See also Thomas Carson:

> The desire-satisfaction theory of welfare needs to be restricted to allow for the logical possibility that a fully informed person could perform acts of self-sacrifice. (Carson, 2000, p. 76)

Thomas Schwartz independently makes more or less the same criticism, claiming that preferentism

> . . . amounts to a form of *psychological egoism* – the doctrine that everyone, at bottom, wants, seeks, or prefers only his own good. If whatever I prefer is good for me, then I prefer only what is good for me. (Schwartz, 1982, p. 199)

L. W. Sumner also puts the charge in terms of egoism:

> . . . the desire theory . . . tells us that if I do what I most want to do . . . then I necessarily do what is best for me. This is a weaker version of psychological egoism . . . [such] that whenever I do choose what I most want then I necessarily maximize my own self-interest. (Sumner, 1996, p. 134)

Sumner cites Amartya Sen, who, writing before Overvold, charges the desire theory with 'definitional egoism' (Sen, 1977, p. 323).[2]

Each of these philosophers is advancing *the argument from self-sacrifice* against preferentism, or the desire theory of welfare. According to this argument, standard, unrestricted desire theories of well-being fail because they have the absurd implication that self-sacrifice is conceptually impossible. The rationale for this claim is, briefly, as follows. Roughly speaking, for an act to count as an act of self-sacrifice, it must be, at least, (i) voluntary, (ii) informed, and (iii) non-optimific for the agent of the act. But, the argument claims, if (i) and (ii) are satisfied, (iii) cannot be, given a standard, unrestricted preferentist theory of welfare. If an act is voluntary, it is the one the agent most wants to do. If it is also informed, then, on the standard desire theories of welfare, it is thereby in the agent's best interest, and so condition (iii) cannot be satisfied. If we act voluntarily and knowingly, we can't help but be egoists and do what is best for us, given the theory's conception of what is best for us.

By an 'unrestricted desire theory' I mean one according to which *all* of one's desires – or all of the desires one would have if one were fully informed or idealized in some way – are relevant to one's well-being. As the passages above illustrate, the argument from self-sacrifice is widely considered to be a serious problem for the desire approach to welfare. If it doesn't merit abandoning the approach entirely, then it at least requires a substantial revision of its basic idea.

In this paper, I attempt to show that, on the contrary, the simplest imaginable, completely unrestricted desire-based theory of welfare is perfectly compatible with the phenomenon of self-sacrifice – so long as the theory takes the right form. It can do this, moreover, without challenging the assumption that any outcome brought about voluntarily and knowingly is, *ipso facto*, the most desired outcome. I further show that this way of formulating the theory is suggested by many typical, rough statements of the theory and, further still, is independently motivated.

Although this simple, unrestricted theory is fully compatible with self-sacrifice, it might seem, as we will later see, that the theory still doesn't allow for *enough* self-sacrifice. That is, there are cases that might intuitively seem self-sacrificial but that the theory seems to imply are not. Therefore, I then go on to attempt to show that this *new argument from self-sacrifice* contains a fallacy, and that once the alleged case of self-sacrifice is modified so that the fallacy is avoided, it is far less intuitive that we have a case of self-sacrifice on our hands. I conclude that, contrary to popular opinion, considerations of self-sacrifice do not pose a problem for preferentism.

These discussions will bring out a crucial, but often overlooked, distinction between two fundamentally different ways of formulating a desire-based theory of welfare. One kind of desire-based theory determines how good an outcome would be for a person by looking to the person's desires *about* the outcome. The best outcome for the person is the one the person wants most (or would want most if he were idealized in some way). But according to an alternative desire-based approach – the one I believe to be more promising – we determine how good an outcome would be for a person by looking to how well satisfied the desires *within* the outcome would be. The best outcome for the person is the one that best satisfies the desires she will have if it comes about.

1. The notion of self-sacrifice

Self-sacrifice has to do with *actions* – actions are the sorts of things that can be self-sacrificial. But theories of welfare have to do with *states of affairs* – outcomes, lives, futures, other parts of lives – and their value for some subject of welfare. Since theories of welfare themselves say nothing about actions, some connecting principles are needed. To begin, we should distinguish between kinds of putative self-sacrifice. Two kinds of sacrifice can immediately be put aside. One such kind occurs when a person forgoes some good *unknowingly* and/or *accidentally*. We do this all the time, even when we are trying to do what is best for ourselves, and it is not what advocates of the argument from self-sacrifice have in mind. Preferentism is compatible with it. Another kind of sacrifice, if we can call it that, occurs when someone intentionally sacrifices a lesser good for himself for a greater good for himself. Such acts might feel sacrificial if they involve delayed gratification, but, in such cases, there is no *net* sacrifice, which is our concern here.

The kind of self-sacrifice relevant here obeys the following principle: an act exemplifies it only if performing the act makes the agent worse off than he otherwise could have been. In other words, an act is an act of self-sacrifice, in the intended sense, only if some alternative to the act would bring about more welfare overall for the agent than it. Let's say that an act is *in the agent's best interest* just in case the life the agent would lead were she to perform it is at least as good for her as the life she would lead were she to perform any alternative to it. We can then state the principle relevant to our purposes here as follows:

> A Principle about Welfare and Self-Sacrifice: An act is an act of self-sacrifice only if the act fails to be in the agent's best interest.

What about a case in which, if someone doesn't do some act that might seem paradigmatically self-sacrificial, such as donate to charity, the person will be so wracked with guilt that he is worse off not doing the generous act? I believe that we should not, on reflection, want to call such an act a literal sacrifice. After all, if the act is in the agent's best interest, there is, by definition, simply nothing that the agent is sacrificing that is not made up for. Note that none of this implies that people who do such acts should not be praised for or be proud of their generosity and their good effects on others, or not be regarded as having fulfilled duties of benevolence. Let me also emphasize that it is my opponent who needs the Principle about Welfare and Self-Sacrifice to be true, so that it can play a role in the argument from self-sacrifice.

I said above that another necessary condition on self-sacrifice is that the agent *know* that the act fails to be in his best interest. We can keep this additional condition in mind, but we will stick with the weaker principle above. It is strong enough to do the work it is supposed to in the argument from self-sacrifice.

© 2011 The Author
Pacific Philosophical Quarterly © 2011 University of Southern California and Blackwell Publishing Ltd.

2. *One kind of preferentist theory of welfare and the argument from self-sacrifice*

According to the crudest version of this kind of theory, the possible lives a person might lead are ranked directly on the basis of the person's actual desires about those lives: one life is better for someone than another just in case he desires it more than, or prefers it to, the other. Call this theory *Life Preferentism*.[3]

The target of Overvold's argument from self-sacrifice, and many of the arguments of the other philosophers discussed above, is a souped-up version of this subjectivist approach. According to the souped-up theory, we look not at the actual desires the person has for each of her various possible lives but instead at her *idealized* desires: typically, the desires she would have if she 'had been fully exposed to available information' and were to keep this information 'firmly and vividly in mind' (Brandt, 1972, p. 682, quoted in Overvold, 1980, pp. 106–107). So the target of the argument from self-sacrifice is a theory like the following:

> Idealized Life Preferentism: One life is better for a subject than another iff the subject ideally prefers it to the other.

Something relevantly like Idealized Life Preferentism, which is Overvold's target, is attributed by Overvold to Brandt. And something very much like it seems to be the theory Sidgwick is discussing (though not endorsing) in the following well-known passage:

> . . . a man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realised in imagination at the present point of time. (Sidgwick, 1907, pp. 111–112)

Rawls endorses a theory like Ideal Life Preferentism in these remarks:

> A person's good is determined by what is for him the most rational long-term plan of life. . . .
>
> Adjusting Sidgwick's notion . . . we can say that the rational plan for a person is the one . . . he would choose with deliberative rationality. It is the plan that would be decided upon as the outcome of careful reflection in which the agent reviewed, in the light of all the relevant facts, what it would be like to carry out these plans and thereby ascertained the course of action that would best realize his more fundamental desires. (Rawls, 1971, pp. 92–93, 417)

Idealized Life Preferentism doesn't sound immediately implausible. The other day I was choosing from among several possible airline flights. I wanted to pick the best flight – the one that was most in my interest. I did

my best to get as much information as I could about each flight – its price, the time it left, the length of any layovers, etc. I gathered all the information, kept each factor in mind, and decided which package I most preferred. This common practice might be seen as an unconscious endorsement of something like Idealized Life Preferentism. And Idealized Life Preferentism avoids the obvious defect of Simple Preferentism: that whatever we happen to prefer most is automatically best for us. Since I might be mistaken, for example, about the actual length of the layover, I might prefer a flight that is actually worse for me. Idealized Life Preferentism respects and explains this obvious fact.

But Idealized Life Preferentism, many philosophers maintain, has an unacceptable consequence: it implies that self-sacrifice is conceptually incoherent. Overvold, for example, invites us to:

> Suppose a man wants more than anything else that his four sons attend a very expensive private college. He is not a rich man. The closest thing to a tangible asset he possesses is a huge life insurance policy. After carefully considering his options, he resolves to kill himself, making it look like an accident. He does so, and four years later his eldest son begins college. Eventually all four sons complete their education and enjoy very happy and rewarding lives. (Overvold, 1980, p. 108)

It will be useful to supply some additional details. Let's suppose that the father had just two possible lives open to him. In one – which we can call 'L1' – the father commits suicide, his sons collect on the insurance, and they go on to 'complete their education and enjoy very happy and rewarding lives'. The father knows that this is what would happen if he were to commit suicide, and he fully and vividly appreciates these facts. In the alternative possible future, which we'll call 'L2', he does not commit suicide. He struggles to fund his sons' educations, but is unable to do so. His sons go on to lead decent enough lives, though, we will suppose, less happy and rewarding than if they had gone to college. Let's suppose that in L2 our protagonist would have suffered from periodic spells of guilt at his failure to send his sons to school, but that he would have learned to overcome these feelings. He would have gone on to lead a long and overall quite satisfying life, a life well worth living by all accounts, if he were to decide against suicide. We must also suppose that, at the time of his decision, the man knows that this is how things would go if he were to decide against suicide, and, further, that he keeps all of this information vividly before his mind when choosing. We are therefore supposing that his actual preferences (at least concerning this decision) are his ideal preferences.

The fully and vividly informed father surveys these two possible lives: the foreshortened L1 where his children thrive and the long and reasonably satisfying L2 where his children do a little less well. He prefers and

chooses L1. It is obvious that the correct way to describe this is as an act of self-sacrifice. The man sacrifices a long and decent life for himself and takes instead a life cut short, all for the sake of his children's welfare.

But Idealized Life Preferentism disagrees. Since the father ideally prefers L1 to its only alternative, Idealized Life Preferentism implies that L1 is best for the father, and therefore that the act that brings about L1 is in the father's best interest. But, as the Principle about Welfare and Self-Sacrifice above states, if an act is in the agent's best interest, it cannot be an act of self-sacrifice. Nothing is being sacrificed in such a case. Therefore, Idealized Life Preferentism implies that the man's act of suicide is not self-sacrificial. But of course it is: by committing suicide, the man sacrifices a fine life for himself in order to send his kids to college. Clearly, the father fares better overall in L2.

Overvold's point isn't just that the theory goes wrong in the case of the sacrificial father. The point of the argument from self-sacrifice is that the theory goes wrong in every putative case of self-sacrifice. If a theory like Idealized Life Preferentism is true, then self-sacrifice is conceptually impossible. The rationale for this claim was outlined above, in the introductory section of the paper. Since my goal is to show compatibility rather than the incompatibility, my focus will be on particular cases of self-sacrifice rather than on the concept itself.

3. *Another kind of desire-based theory of welfare*

What Life Preferentism and Idealized Life Preferentism (and, incidentally, Overvold's own restricted desire-theoretic proposal for avoiding the argument from self-sacrifice) have in common is the manner in which they rank lives in terms of welfare. Metaphorically speaking, they lay out before the subject the whole lives the subject might lead; then they ask the *subject* to rank the lives: Which life do *you*, dear subject, like best? This is *subjectivism par excellence*. It also seems to be the feature that makes the theory go wrong with respect to self-sacrifice. A better approach, in my view, looks not for which whole life the subject has the strongest desire, but for which life is such that all of the desires the subject would have throughout that life would be best satisfied. Don't, theory of welfare, ask the subject which life she prefers. Instead, examine the lives yourself; take note of each occasion in each life in which the subject gets what she wants and each occasion in which she doesn't get what she wants. Whichever life contains the greatest balance of the former over the latter is the best life. Never mind which life the subject happens to prefer right now. Since this approach, in ranking lives, de-emphasizes how the subject feels about those lives, it is in a way less subjectivistic.

We can state this approach as follows:

> Desire Satisfactionism: One life is better for a subject than another iff it contains a greater balance of desire satisfaction over frustration than the other.

The Desire Satisfactionist picture is this. Each day of our lives, and over greater spans as well, we get some things we want and fail to get other things we want. When the former happens, we enjoy a *desire satisfaction*; when the latter happens, we suffer a *desire frustration*. The greater the number and intensity of the satisfactions and the lesser the number and intensity of the frustrations, the greater the balance of desire satisfaction over frustration.[4]

Some of our desire satisfactions involve *global desires*: we want our lives to go a certain way. Others are *local*: we want our team to win tonight. Some are *self-regarding*: we want fame or recognition. Others are *other-regarding*: we want our kids to have good lives (or, again, our team to win). Desire Satisfactionism is *unrestricted*: it includes all such desires. It is important that an unrestricted theory is, as I try to demonstrate in a moment, compatible with self-sacrifice because, as you would expect, having to restrict a desire theory to count only a subset of our desires in order to avoid an objection often brings with it new problems.

This, incidentally, is what I think is wrong with Overvold's own theory, which he proposes as a desire-theoretic way to avoid his argument from self-sacrifice. Overvold's view is restricted to count as relevant to one's well-being only desires that, roughly speaking, involve oneself. But it is counterintuitive, and also anathema to the spirit of the preferentist approach to well-being, to ignore some non-self-regarding desires, such as those for the success of one's children or one's team, or those having to do with one's most cherished interests, projects, or goals. If tonight my heart's desire is that my team win, and then they do in fact win, how can any theory of welfare – let alone a desire theory of welfare – deny that things went well for me tonight, in at least this respect?

Desire Satisfactionism is *actualist*: it counts the desires one actually has. Its *idealist* cousin is:

> Idealized Desire Satisfactionism: One life is better for a subject than another iff it contains a greater balance of ideal desire satisfaction over frustration than the other,

where an ideal desire satisfaction or frustration is the satisfaction or frustration of an ideal desire – a desire one would have if one were fully and vividly informed.

It is worth mentioning that the feature that, as we'll see, enables this theory to handle self-sacrifice – i.e. formulating the desire theory in this

slightly more objectivistic way – also renders the move to idealize less urgent. Suppose I want to go cycling, not knowing it will rain, and that I would have chosen to stay in if I knew about the rain. Although, in my ignorant state, cycling is what I now most want to do, Desire Satisfactionism does not imply, as Life Preferentism does, that it is therefore in my best interest to go cycling. Since it will rain, the life I would lead were I to go cycling contains a lesser balance of desire satisfaction over frustration – due to how strongly I will want not to be cycling in the rain once I am stuck in it – than the life I would lead were I to stay home – even while wishing it weren't raining so that I could be cycling. Life Preferentism needs to idealize in order to handle this case properly, but Desire Satisfactionism does not.

The differences between Desire Satisfactionism and Idealized Desire Satisfactionism won't matter in what follows. The cases to be considered all involve idealized agents, about whom non-idealized and idealized theories agree. I will therefore sometimes just mention Desire Satisfactionism, although what I say will apply to both theories.

Simple and Ideal Desire Satisfactionism *are* subjectivist in perhaps the most important sense: according to them, welfare has to do with our attitudes towards what befalls us in life rather than with the things themselves that befall us. But the theories are objectivistic with respect to the evaluation of whole lives: the subject's attitudes about her life as a whole play no special role in determining how good it is. This is, for our purposes, the crucial way in which the Desire Satisfactionist theories differ from the Life Preferentist theories.[5]

To appreciate further the differences between the Desire Satisfactionist and Life Preferentist approaches, note that the Life Preferentist theories appeal only to one's *present* desires (whether actual or ideal) to determine whether some act is in one's best interest. On Desire Satisfactionism, by contrast, one's *future* desires are directly relevant, too. Furthermore, one's present desires have no special status in determining how good an outcome would be on this approach. They count just as much as any (equally intense) future desire.

Desire Satisfactionism is probably the simplest possible theory of this sort – 'square one' for this form of desire-based theory. I do not claim that it stands in no need of refinement.[6] My point is rather that even the simplest possible, totally unrestricted sort of desire theory – so long as it takes this form – easily accommodates the phenomenon of self-sacrifice.

4. *Desire satisfactionism and the possibility of self-sacrifice*

Desire Satisfactionism straightforwardly handles the phenomenon of self-sacrifice. Given Desire Satisfactionism, an act is in its agent's best interest iff no alternative to it would produce a greater balance of desire satisfaction

over desire frustration for the agent. An act can be in the agent's best interest if it leads to the satisfaction of a large enough number of intense enough desires. Or it can do so if its alternatives would lead to enough frustration.

The sacrificial father faces a choice between L1, the suicide life, and L2, the longer life. We can suppose that L1 contains a few desire frustrations in its last moments – those that naturally accompany impending death. L1 also contains some desire satisfactions not present in L2, most notably the desire that the sons attend college. The relevant features of L2 are far too numerous to state, since the life is long. But recall the general course of the life: the man's frustrated desire that his sons attend college continues, but eventually he 'gets over it' and goes on to lead a fine life.

Given these descriptions of L1 and L2, it is clear that, according to Desire Satisfactionism, nothing bars L2 from having a far greater value for the father than L1 – this despite the fact that the father prefers L1 to L2. L2 is so much longer that we are free to describe it as having any number of intense desire satisfactions. True, L2 will contain recurring frustrations over the failure of the sons to attend college, but there is nothing to stop us from stipulating that these frustrations are heavily outweighed. So let L2 be such that its balance of desire satisfaction over frustration is far greater than the overall balance in L1. According to Desire Satisfactionism, then, L2 is better for the man than L1. The theory thus opens the door for self-sacrifice, for, according to the theory, a person, even a fully and vividly informed person, can prefer and choose what is worse for him for the sake of someone or something he cares about.

4.1. AN OBJECTION

Perhaps the advocates of the argument from self-sacrifice would attempt to block this line of reasoning by appeal to a principle like the following:

> A Principle about Preference: If a subject knows that some outcome O1 contains a greater balance of desire satisfaction over frustration for the subject than does an alternative outcome O2, and the subject vividly appreciates these facts, then the subject prefers O1 to O2.

The idea behind A Principle about Preference may be attractive enough. Suppose I have complete and vivid knowledge about two alternative lives of mine, + and –. + satisfies many of its desires and frustrates none, while – frustrates many and satisfies none. How could I possibly prefer + to –? What features would attract me to –? None, it would seem, for if there were any, then – would have to satisfy my desires for them.

Despite its initial plausibility, A Principle about Preference does not stand up to scrutiny. The root of the problem is that it is possible for a person to know, even vividly, that he will desire certain things in the

future, and yet fail to be moved in the present to behave in such a way that those future desires will be satisfied. I might prefer today the alternative of not visiting the dentist even though I know vividly that visiting the dentist would better satisfy my desires in the long run. This may make me irrational or weak-willed, but that doesn't matter, for such irrationality is not always curable by additional knowledge and vivid appreciation. And, in any event, not only the weak-willed falsify the principle. Imagine an evil genius who gives a good Samaritan two options:

Option A: a life of decadence with desires for decadence; misery for everyone else without desires to the contrary;
Option B: a life of toil with desires to the contrary; flourishing for everyone else without desires that others flourish.

The good Samaritan sees that Option A would give herself a far greater balance of desire satisfaction over frustration in her life. But being a benevolent Samaritan, she cannot bring herself to choose it. She prefers Option B. So A Principle about Preference, which implies that she will not prefer Option B, is false.

5. *Three further points about this theory*

5.1. THE STANDARDNESS OF DESIRE SATISFACTIONISM

There are three further points worth making about this slightly less subjectivistic, Desire Satisfactionist way of formulating desire theories of welfare. The first is that, although the distinction between the two ways of ranking that I have been discussing is often glossed over, many rough statements of desire theories of welfare are at least as suggestive of the second, more objectivistic sort of way. For example, in discussing 'Desire-Fulfillment Theories,' Derek Parfit describes the simplest such theories as those that claim

> . . . that what is best for someone is what would best fulfill *all* of his desires, throughout his life. (Parfit, 1984, p. 494)

This is suggestive of including future desires, as Desire Satisfactionism does. Richard Kraut's characterization of the view that 'equate[s] the human good with the satisfaction of desire' is similarly implicative:

> Roughly speaking, what makes a state of affairs good for someone is its satisfaction of one of that person's desires; accordingly, our lives go well to the extent that our desires, or the ones to which we give the greatest weight, are satisfied. (Kraut, 1994, p. 164)

And according a recent characterization by Fred Feldman,

> Preferentists maintain that what makes a life good is that desires of some sort are satisfied rather than frustrated *within* that life. (Feldman, 2004, p. 16, emphasis added)

It therefore seems mistaken to suggest, as many writers do, that *standard* versions of the desire theory leave no room for self-sacrifice. The desire satisfactionist way of formulating the theory needn't be seen at all as a *response* to the argument from self-sacrifice – a retreat to a non-standard form of the view. It is the sort of formulation philosophers often have in mind when stating what they take to be the standard version of the theory, and it is a formulation that is consistent with most all rough statements of the theory.[7]

5.2. AN INDEPENDENT MOTIVATION FOR THE MORE OBJECTIVISTIC FORMULATION

The second point worth making also supports the idea that formulating desire theories in the more objectivistic way is by no means an *ad hoc* response to the argument from self-sacrifice. It is independently motivated by the need to avoid a fundamental defect in the first way of formulating desire theories.

The defect has to do with time. Theories of welfare are supposed to deliver, among other things, verdicts about the overall welfare value of lives. The first, more subjectivistic sort of desire theory does this by consulting the subject's attitudes, or idealized attitudes, about the lives she could lead. But we have attitudes, even idealized attitudes, at times. And the attitudes one has about one and the same life, even the idealized attitudes one has about it, can change over time.[8] Thus, to avoid the result that a single life can be both an overall good life and an overall bad life for the subject, these theories must do something unacceptable, even unintelligible: they must *relativize the overall welfare value of lives to times*. But just as it makes no sense to ask, say, *how long* some life was relative to some time, it makes no sense to ask *how good overall* some life was relative to some time.[9]

Desire Satisfactionism faces no such problem. The overall balance of desire satisfaction over frustration (ideal or otherwise) contained in some whole life is, like the overall amount of well-being in a whole life, an unchanging quantity.

The subjectivistic way of evaluating lives might be able to make its relevant quantity unchanging by *averaging over time* the attitudes the subject has to her whole life. Such an approach is suggested by Bricker (1980) and by Carson (2000, p. 86). But I think such a view is properly

characterized as of the more objectivistic sort (on the present taxonomy), since, according to it, how good some life one could now choose to lead is not determined exclusively by the attitudes, idealized or otherwise, one now has towards that life.

5.3. AN INDEPENDENT MOTIVATION FOR THE MORE SUBJECTIVISTIC FORMULATION?

The third point concerns the internalism/externalism debate about reasons, motivations, and value, considerations of which might seem to provide an independent reason for preferring the subjectivistic, Life Preferentist way of formulating the desire theory. One alleged attraction of the whole desire approach to well-being is that it seems well-suited to obey an 'internalism requirement,' according to which, on one way to put it,

> . . . it is a necessary condition on something being good for a person that she be capable of caring about it. (Rosati, 1995, p. 300, 10n)[10]

Robert Noggle even sees its amenability to internalism as 'the intuition behind desire-based accounts of well-being,' describing internalism as follows:

> The fact that desire-based theories make our well-being something that matters to us seems to be an advantage over theories that simply posit a list of things that make a person's life go well, whether they matter to the person or not. . . . The reason that such internalism seems appropriate in a theory of well-being is that if we are measuring the extent to which a life is valuable for the person living it, then it seems that the criteria for evaluation must be those of the agent herself. The ends and goals in terms of which we evaluate the success of a person's life must not, it seems, be completely alien to the agent's own ends and goals. (Noggle, 1999, p. 303)

The theory that most straightforwardly satisfies the internalism requirement (on Rosati's statement of it) is Life Preferentism. For on Life Preferentism, the life that is best for a person isn't just one she is capable of caring about, it is the one she actually cares most to lead. It isn't quite as obvious that Idealized Life Preferentism satisfies the requirement, since it needs to be decided how best to understand '*capable* of caring'. But one natural way to understand it is as meaning that the subject would care about the putative good in question if she were fully and vividly informed about it. If this is how the internalism requirement is understood, then Idealized Life Preferentism satisfies it (perhaps by design). But the Desire Satisfactionist theories might seem to run afoul of the internalism requirement, since on these views how good some outcome would be for someone is not determined by her present desires about it, and so need not connect

up in any way to what she presently cares about, or is presently capable of caring about.

But it is not at all clear that our internalist intuitions about welfare – and even the Rosati formulation above – excludes future carings. The idea that it is a necessary condition on something being good for a person that she be capable of caring about it, *either now or in the future when she gets it*, is no less intuitive than a variant of that idea restricted to the present. Similarly, when Noggle reports the idea that if we are measuring the extent to which a life is valuable for the person living it, the criteria for evaluation must be those of the agent herself, this sounds plausible, but, again, Desire Satisfactionism can meet it. Desire Satisfactionism doesn't impose an objective list of goods – as do the theories Noggle sees as most firmly in opposition with the internalist requirement – but looks only to the agent's own desires. On Desire Satisfactionism, the ends and goals in terms of which we evaluate the success of a person's life are by no means completely alien to the agent's own ends and goals, since they *are* the agent's own ends and goals – albeit some of them merely future ends and goals.

Ideal Desire Satisfactionism's ability to satisfy the internalist requirement isn't quite as certain. But it is no less certain than Idealized Life Preferentism's ability (they would seem to rise or fall together on this front). Thus the objectivistic and the subjectivistic forms of desire theory fare the same here: the simple version of each pretty clearly satisfies the requirement, and the ideal version of each may satisfy it too.[11]

Some philosophers, however, are especially concerned to connect one's good (and other reason-giving phenomena) to one's *present* carings and motivations.[12] Non-idealized Desire Satisfactionism and Idealized Desire Satisfactionism, unlike non-idealized and Idealized Life Preferentism, can deliver no such connection. But the claim that one's good is necessarily connected only to what one is presently capable of caring about is, in fact, a far more controversial thesis, and a far cry from the relatively innocuous idea that our good must be connected to our own ends and goals rather than imposed by an objective list. Indeed, it seems to me that, pretheoretically, the stronger internalist requirement – the one restricted to one's present carings – is in fact less plausible than its denial. For it is a sad commonplace that some people cannot be made to care about their own future welfare.

I conclude that non-idealized and Idealized Desire Satisfactionism are, respectively, as friendly as non-idealized and Idealized Life Preferentism are to the plausible version of the internalist requirement. The subjectivistic way of formulating a desire-based theory of welfare therefore receives no independent support from considerations of internalism.

6. *A new argument from self-sacrifice*

I concluded earlier that Adams, Brandt, Carson, Darwall, Griffin, Haslett, Overvold, Schwartz, Sen, and Sumner – advocates of the argument from self-sacrifice – are mistaken when they claim that the desire theory of welfare cannot recognize the possibility of self-sacrifice, or that it can only if the set of desires relevant to welfare is restricted in some way. However, some might think that Desire Satisfactionism still fails to allow for enough self-sacrifice. The case above of the suicidal father is 'highly diachronic': self-sacrifice occurs because the father chooses today to forgo goods he would receive in the future, including the very distant future. This is one of the features of the case that enables Desire Satisfactionism, a theory that counts future desires, to accommodate it. But consider a 'more synchronic' case, a case in which someone chooses now to forgo goods she would be receiving now:

> *Alice's Friday Night*: Alice is deliberating over how to spend her Friday night. She can go to the disco with her friends, or she can volunteer at the soup kitchen. Alice considers the options and, despite how badly she wants to go dancing with her friends, she decides, voluntarily and with full and vivid knowledge, to spend her Friday night helping the needy at the soup kitchen. She feels it would be the right thing to do, and so she does it.

The proponent of the argument from self-sacrifice then asks us to consider how we should explain Alice's decisions to go, and continually remain, at the soup kitchen. One sort of explanation goes as follows. Alice knowingly and voluntarily chooses to spend her Friday night at the soup kitchen, and continually re-affirms that choice by remaining there. Thus, this must be the option she most prefers throughout the evening. That is, it must be that, despite what it's like to be at the soup kitchen and what it would be like, as Alice vividly appreciates, to be at the disco, Alice nevertheless desires to be at the soup kitchen more than she desires to be at the disco. How else to explain why she stays? It would then seem to follow that she gets more desire satisfaction from being at the soup kitchen. And if that's true, then Desire Satisfactionism implies that her act is not an act of self-sacrifice, since, assuming that the soup kitchen and the disco are her only two alternatives (and that the rest of her life will be the same no matter what she does on this Friday night), Desire Satisfactionism implies that she is doing what is best for her.

This, the argument based on Alice's Friday Night, we can call the *new argument from self-sacrifice*. As with the original argument from self-sacrifice, it makes use of the Humean idea that an outcome a person voluntarily and knowingly chooses is the outcome the person wants most. This claim is questionable – it overlooks the Kantian idea that beliefs

(specifically, our evaluative beliefs) might motivate in addition to, and in opposition to, our desires. Rejecting the Humean view might provide an interesting and plausible rebuttal to the new argument from self-sacrifice. If evaluative beliefs can motivate and explain action, a natural thing to say about Alice is that she goes to the soup kitchen not because she particularly wants to be there, but because she believes she ought to be there. Since Alice wouldn't then be getting much desire satisfaction from being at the soup kitchen, this would be, given Desire Satisfactionism, a worse option from the standpoint of her welfare. It would then be a sacrifice for her to be there, and the new argument from self-sacrifice would fail.

This strategy would, however, leave the desire-based theory of welfare a hostage to fortune: if the Kantian theory of motivation turns out to be false, this strategy is undermined. Fortunately for preferentists, they need not rely on this option, for the reasoning above fails in a different way. The other main inference in the reasoning above moves from the claim

(i) that Alice desires to be at the soup kitchen more strongly than she desires to be at the disco,

to the claim

(ii) that Alice receives more desire satisfaction from being at the soup kitchen than she would receive from being at the disco.

But this inference is fallacious. It assumes that, in order to measure the balance of desire satisfaction over frustration that Alice would have received had she gone to the disco, we appeal to the desires that she actually has at the soup kitchen toward the prospect of being at the disco and what would befall her there. But this is a misunderstanding of Desire Satisfactionism. Desire Satisfactionism determines how good a counterfactual scenario would have been for a person by looking at the desires the person has *in the counterfactual scenario* – not at the desires the person *actually* has towards what would befall her in the counterfactual scenario. These sets of desires need not be the same, and often are not the same, even for idealized desirers.

Note that this is somewhat analogous to the way Desire Satisfactionism treats the future. It determines how good some future would be for a person by looking at the desires the person has *in that future* – not just at the desires the person has *now* towards what befalls her in that future.[13]

To illustrate, we can note that being at the disco would very likely give rise in Alice to new desires – desires to be dancing with the person with whom she is dancing, desires to be hearing the song she is hearing, desires to be sipping the drink she is sipping, etc. It would also be likely to cause an increase in the intensity of some of the desires she also has at the soup

kitchen; while at the soup kitchen, she does have at least some desire to be at the disco, and it is not unlikely that being at the disco would make her want to be there even more. Nor is it unlikely that being at the disco would cause her desires to be at the soup kitchen to wane. Since she is an ideal agent, Alice would, while at the soup kitchen, vividly know about these facts – about the experiences she'd be having at the disco, and her desires for them – but this need not translate into equally strong desires for these experiences while she is at the soup kitchen. Indeed, it better not. Otherwise such desires would (assuming they are sufficiently strong) motivate her to leave for the disco, and the case would not work as the new argument from self-sacrifice requires (it requires her to spend her Friday night at the soup kitchen). So it must instead be that a person can vividly know about the experiences she could be having and her desires for them were she doing something else, but still have her actual desires for those experiences remain less strong. This is analogous to what we learned in our study of the original argument from self-sacrifice – viz., that a person can vividly know about the experiences she would be having in the future and her desires for them, without that knowledge translating into a motivation to bring about those experiences.

To try to get the case to work properly, the proponent of the new argument from self-sacrifice must then stipulate that, had Alice gone to the disco instead, this wouldn't have affected her desires in such ways (or at least not enough to make this alternative contain a greater amount of desire satisfaction than the soup kitchen alternative; in order for the case to work as intended, Desire Satisfactionism cannot imply that Alice would have had a better night at the disco). It must be stipulated that, had Alice gone to the disco, she would have continued, throughout her whole time at the disco, to have fairly strong desires to be at the soup kitchen, and not to have formed many desires, or many very strong ones, for what befell her at the disco. The proponent of the new argument from self-sacrifice must, that is, just stipulate claim (ii) above, the claim that Alice receives more desire satisfaction from being at the soup kitchen than she would have received from being at the disco.

But this stipulation – which the new arguer from self-sacrifice is forced to make due to the invalidity of the inference above – seems to undermine the argument. For now the facts of Alice's Friday Night are as follows:

- In spending her evening at the soup kitchen, Alice is getting what she most wants on this evening;
- Alice will not lose out in the future on things she will be wanting in the future by going to the soup kitchen tonight (contrast this with Overvold's sacrificial father);
- Had Alice gone to the disco instead, she would have, during her whole time there, been fairly strongly wanting to be at the soup kitchen.

- Had Alice gone to the disco instead, she would not have formed all manner of new desires for what befell her at the disco, or had a very strong desire to be there at the disco.[14]

Because all this is true, it is no longer very intuitive that Alice is doing what is worse for her on this Friday night. Indeed, the claim appears rather question-begging. Just because Alice isn't acting with *herself* in mind when she goes to the soup kitchen, or with her own best interests in mind, and is instead acting benevolently for others, we cannot conclude that she must therefore fail to be doing what is in her best interests.[15] As the paradox of hedonism has taught us, we are often more likely do what is best for ourselves when we are not trying to do what is best for ourselves. Furthermore, compare the case of Alice to Overvold's case of the sacrificial father. Our intuition that the father isn't doing what is best for himself is very clear and strong. It would not have been at all plausible for a desire-theorist to reject it. The only way out was to show, as I tried to do, that the desire theory is actually fully compatible with the intuition. But the analogous claim about Alice's Friday Night is not comparably intuitive. And, finally, when we remember that Alice's act need not count as self-sacrificial in order for it to count as admirable and worthy of our praise, I believe that any temptation to insist that her act is self-sacrificial dissolves.

* * * * *

Even though no self-sacrifice argument undermines either non-idealized or Idealized Desire Satisfactionism, it is not here claimed that either of these theories is fully adequate as it stands. There are many objections to desire-based theories of welfare worthy of attention. But one such line of objection, the argument from self-sacrifice, seems, despite its popularity, not to threaten the idea that welfare is fundamentally a matter of getting what one wants.[16]

Department of Philosophy
University of Colorado at Boulder

NOTES

[1] See, e.g., Brandt, 1966 and Brandt, 1972.

[2] To leave no doubt as to the popularity of this objection, here are three more examples. Robert Adams claims that unrestricted desire-satisfaction theories face

> . . . the paradox that if under the relevant conditions I would choose, all things considered, for the sake of my ideals, or for the good of others, to sacrifice my own comfort, tranquility, physical and social pleasures, health of mind and body, and length of days, then that is what is *best for me*. (Adams, 1999, p. 89)

D.W. Haslett writes,

. . . the preference model seems to allow *too much* to count as personal welfare. . . . by allowing too much to count as personal welfare, this model precludes us from making certain important distinctions between personal welfare and what is not personal welfare. One such distinction precluded by this model is the distinction between personal welfare and self-sacrifice. Say, for example, a person prefers above all else to set himself on fire as a protest against nuclear weapons, thinking that, although it will regrettably terminate his life, such a dramatic protest might be of immense benefit to humanity. Obviously, for him, to burn himself to death for the benefit of humanity would be an act of great self-sacrifice. But because it would be satisfying his strongest preference, the preference model commits us to saying that burning himself to death is actually in his own interests. (Haslett, 1990, pp. 79–80)

And Brad Hooker (1996, p. 144) cites Overvold's argument approvingly as well. Although Hooker pretty clearly accepts Overvold's argument, he nevertheless alludes, a paragraph later, to something like the sort of solution I will be proposing here.

[3] Since, I assume, preference and desire are interdefinable, there is no substantive difference between theories of welfare stated in terms of desire and those stated in terms of preference.

[4] We are not talking here about *feelings* of satisfaction and frustration. A desire is satisfied (frustrated) just in case the thing the subject wants to come about in fact comes about (fails to come about). Sometimes, of course, the desired things are feelings of satisfaction.

[5] That the Desire Satisfactionist theories appeal to desire and the Life Preferentist theories to preference is not a substantive difference between the theories. I choose these different labels mainly for ease of distinguishing them.

[6] My own view about how this simple theory should be improved is described in Heathwood, 2006 and laid out and defended much more fully in Heathwood, unpublished.

[7] Many characterizations of the desire approach seem to be ambiguous between the two different kinds of formulation. In each of the following passages, it is unclear whether the author means to include only present desires, or both present and future desires:

The simplest form of desire account says that utility is the fulfilment of actual desires. (Griffin, 1986, p. 10, emphasis removed)

According to the preference theory, well-being consists in having one's preferences satisfied. To the extent that your preferences or desires are satisfied, you are better off; to the extent that your preferences or desires are not satisfied, you are less well-off. (Kagan, 1994, p. 312)

. . . the desire-satisfaction view of welfare holds that people's well-being consists in their desires being satisfied. (Shaw, 1999, p. 53)

One prominent nonhedonist account of welfare . . . is the *desire fulfillment theory of welfare*. Its basic idea is that what makes a person's life go better is the fulfillment of her desires, and what makes it go worse is the nonfulfillment of her desires. (Timmons, 2002, p. 143)

It should be noted that some recent discussions, e.g. Crisp, 2006, are sensitive to the importance of our distinction.

[8] Some philosophers, e.g. Carson (2000, p. 85), speculate that one's idealized preferences might be unchanging. But, so long as the idealization is defined in a non-welfare-value-laden way, as it must be to avoid circularity, I don't see how there could be a necessary connection between the distinct existences of an idealized preference and its object.

[9] I assume that those who are concerned about special relativity also know how to reformulate this argument to avoid these concerns.

[10] Rosati is here describing the 'existence internalism' of Darwall (1983, pp. 54–55).

[11] If anything, this amounts to an advantage for the Desire Satisfactionist approach, since the non-idealized version of it has more promise than the non-idealized version of the Life Preferentist approach (as discussed briefly above, in Section 3).

[12] The most well-known example may be Williams (1981).

[13] For Desire Satisfactionist theories restricted so as to ignore what R. M. Hare (1981, p. 101) calls 'now-for-then' desires, the way they treat the future will be *exactly* analogous (not just somewhat analogous, as I say above) to the way they treat counterfactual scenarios.

[14] The basic idea is that Alice's Friday Night cannot have Alice being such that, had she gone to the disco, she would have had very many, very strong disco-favoring desires and soup-kitchen-disfavoring desires. For to the extent that this is true, the disco will be to that extent a good option for Alice even on Desire Satisfactionism, and so will to that extent undermine my opponent's claim that Alice's act is not self-sacrificial given Desire Satisfactionism.

[15] Maybe Schwartz, in his discussion of the argument from self-sacrifice, commits (or comes close to committing) something like this mistake when discussing how only self-regarding preferences would be tied to welfare. He writes:

Roughly speaking, self-regarding preferences are ones not based on any ultimate objective of promoting the welfare, the goals, or the happiness of anyone but their subject. Only such preferences (and perhaps not even they) constitute strong evidence of what is good for their subject. (Schwartz, 1982, p. 199)

[16] A distant ancestor of this paper was presented at the 2004 Central Division Meeting of the American Philosophical Association. I thank my commentator there, Mark van Roojen, as well as members of the audience for helpful feedback. I would also like to thank David Barnett, Donald Bruckner, Dan Doviak, Kris McDaniel, Doug Portmore, Jason Raibley, several anonymous reviewers, and especially Fred Feldman for helping me improve the paper.

REFERENCES

Adams, R. M. (1999). *Finite and Infinite Goods*. Oxford: Oxford University Press.

Brandt, R. B. (1966). 'The Concept of Welfare,' in S. R. Krupp (ed.) *The Structure of Economic Science*. Englewood Cliffs, NJ: Prentice-Hall.

Brandt, R. B. (1972). 'Rationality, Egoism, and Morality,' *Journal of Philosophy* 69, pp. 681–697.

Brandt, R. B. (1982). 'Two Concept of Utility,' in H. B. Miller and W. H. Williams (eds) *The Limits of Utilitarianism*. Minneapolis, MN: University of Minnesota Press.

Bricker, P. (1980). 'Prudence,' *Journal of Philosophy* 77, pp. 381–401.

Carson, T. L. (2000). *Value and the Good Life*. Notre Dame, IN: University of Notre Dame Press.

Crisp, R. (2006). 'Well-Being,' in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2006 Edition)*, URL = <http://plato.stanford.edu/archives/win2006/entries/well-being/>.

Darwall, S. (1983). *Impartial Reason*. Ithaca, NY: Cornell University Press.

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

Feldman, F. (2004). *Pleasure and the Good Life*. Oxford: Oxford University Press.

Griffin, J. (1986). *Well-Being*. Oxford: Oxford University Press.

Hare, R. M. (1981). *Moral Thinking*. Oxford: Oxford University Press.

Haslett, D. W. (1990). 'What is Utility?' *Economics and Philosophy* 6, pp. 65–94.

Heathwood, C. (2006). 'Desire Satisfactionism and Hedonism,' *Philosophical Studies* 128, pp. 539–563.
Heathwood, C. (unpublished) 'Subjective Desire Satisfactionism'.
Hooker, B. (1996). 'Is Moral Virtue a Benefit to the Agent?' in R. Crisp (ed.) *How Should One Live?* Oxford: Clarendon Press.
Kagan, S. (1994). 'Me and My Life,' *Proceedings of the Aristotelian Society* 94, pp. 309–324.
Kraut, R. (1994). 'Desire and the Human Good,' *Proceedings and Addresses of the American Philosophical Association* 68, pp. 39–54.
Noggle, R. (1999). 'Integrity, the Self, and Desire-Based Accounts of the Good,' *Philosophical Studies* 96, pp. 303–331.
Overvold, M. C. (1980). 'Self-Interest and the Concept of Self-Sacrifice,' *Canadian Journal of Philosophy* 10. pp. 105–118.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Belknap.
Rosati C. S. (1995). 'Persons, Perspectives, and Full Information Accounts of the Good,' *Ethics* 105, pp. 296–325.
Schwartz, T. (1982). 'Human Welfare: What It Is Not,' in H. B. Miller and W. H. Williams (eds) *The Limits of Utilitarianism*. Minneapolis, MN: University of Minnesota Press.
Sen, A. K. (1977). 'Rational Fools: A Critique of the Behavioral Foundations of Economic Theory,' *Philosophy and Public Affairs* 6, pp. 317–344.
Shaw, W. H. (1999). *Contemporary Ethics*. Oxford: Blackwell.
Sidgwick, H. (1907). *The Methods of Ethics*, 7th edn. London: Macmillan and Company, Ltd.
Sumner, L. W. (1996). *Welfare, Happiness, & Ethics*. Oxford: Clarendon Press.
Timmons, M. (2002). *Moral Theory*. Lanham, MD: Rowman & Littlefield.
Williams, B. (1981). 'Internal and External Reasons,' in B. Williams (ed.) *Moral Luck*. Cambridge: Cambridge University Press.