Well-being, Self-Sacrifice, and Core Desires

Picture a burning building surrounded by firefighters and a crowd of onlookers. The fire chief has called his men back, because it has become too dangerous to go inside. Suddenly they see children at a window on a high floor, trapped inside. Michelle, one of the onlookers, runs into the building, ignoring the warning cries of the firefighters. She gets the children out safely but suffers serious and lasting injuries herself.

The case thus provides an example for one of the most serious difficulties faced by desire-based theories of well-being\(^1\), which hold that all that is good for a person is to have her desires satisfied, and all that is bad for her is to have them frustrated. It is widely believed that such views cannot make sense of cases of self-sacrifice.

This paper aims to show that overcoming this problem requires a new way of constructing desire-based theories of welfare, and propose a particular way to do so. To this end, I will first argue that the most promising approach to solve the problem – put forward by Chris Heathwood\(^2\) - does not quite do the trick. An analysis of this failure reveals why no extant version of the desire view can succeed. This leads to my proposal for an original desire-based theory of well-being. I will try to motivate the proposal and show how it can succeed where others could not.

I

Desire-satisfaction based theories of well-being (desire theories, for short) hold that a person’s well-being is solely constituted by the satisfactions of that person’s desires.\(^3\) What’s good

---

\(^1\) I am using “well-being” and “welfare” interchangeably.

\(^2\) Heathwood (2011).

\(^3\) Where I speak of desire satisfaction, others prefer talk of fulfillment of preferences or wants. Nothing should hang on this for my arguments in this paper.
for a person is for her to get what she most wants. Since Michelle wants most to save the children, doing so satisfies her strongest desire, outweighing desires to live on unharmed by fire. On standard desire theories, her sacrificing herself to save the children therefore increases her well-being. Desire theories seem to make self-sacrifice “conceptually impossible”. Many have taken this to be a serious, even decisive difficulty for desire theories. Why is it so important that a theory of welfare can make sense of this concept?

Because there is a difference between someone joining the resistance against an oppressive regime, and someone joining a white shoe law firm. Both are doing what they most want to do, but we are impressed with the freedom fighter, in a way we are not with the lawyer. We do not usually think of lawyers as sacrificing a chance to risk their life. Those who risk their life for good causes seems worthy of special praise. There is something impressive about people willing to sacrifice themselves for others. We might feel reverence for them. A theory that cannot make sense of self-sacrifice cannot make sense of that reverence. It treats the lawyer and the freedom fighter are equal.

II.

Until recently, no convincing solution to the problem from self-sacrifice had been put forward. The standard move to save desire views from problematic implications is to incorporate an element of idealization (e.g. by allowing only ideally informed or rational desires to count towards well-

---

4 In his influential discussion of this problem, Mark Overvold uses the case of a father who sacrifices his life to secure a good education for his sons. See Overvold (1980), p. 108.
6 See e.g. Brandt (1982); Schwartz (1982); Griffin (1986); Kraut (1994); Adams (1999); Arneson (1999); Darwall (2002).
7 This need not necessarily be moral praise. People can sacrifice themselves for evil causes.
being). A particularly sophisticated suggestion was put forward by Peter Railton, according to whom what is good for me is the satisfaction of those of my desires which an idealized version of myself would want me to want to satisfy. But who is to say that Michelle’s idealized self wouldn’t want her to want to save the children from the burning building? Similarly for less sophisticated idealized theories: even an agent with idealized desires could still want to sacrifice herself. None of the usual ways to exclude the problematic desires seemed to work. Heathwood’s ingenious proposal finally seems to do the trick. He proposes that “an act is an act of self-sacrifice only if the act fails to be in the agent’s best interest,” where an act counts as being in an 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.”

Now he can bring in his version of desire theory, which I will call HDS (for “Heathwood’s Desire Satisfactionism”). According to HDS, “one life is better for a subject than another iff it contains a greater balance of desire satisfaction over frustration than the other.” To determine whether the act in question is in fact bad for the agent (which self-sacrifice is supposed to be) HDS considers the entirety of the agent’s remaining life after the act, summing up all of the desire satisfactions and frustrations it would contain. But, as Heathwood points out, our desires change over time, in ways we often fail to anticipate. In determining which life will be best, HDS includes satisfactions and frustrations of desires the agent would have, rather than only those she has at the time of the act. This is the crucial move.

---

10 For a brief overview of the debate, see Heathwood (2011), pp. 18-19.
11 Heathwood (2011), p. 21
12 This is not supposed to be his own preferred theory, which he defends in his (2006), although there is overlap.
Michelle’s desire to save the children outweighs all the desires she then has about her potential future life, but not all the desires she would come to have in the future, were she to live on.\textsuperscript{14} Living out her life would therefore yield a higher net sum of satisfaction over frustration than her cut-off life of self-sacrifice. According to HDS, Michelle’s choice is in fact bad for her, avoiding the counter-intuitive result of standard desire theories. In my view, HDS is the strongest contender for a solution to the problem of self-sacrifice.\textsuperscript{15}

III.

However, in certain kinds of cases – call them \emph{Happy Ending Cases} – HDS fails by its own lights. Consider:

\textbf{Michelle’s rise to fame}

Suppose after many weeks in the hospital, Michelle is interviewed by a major talk show about her self-sacrificial act. She becomes nationally known and receives an attractive and lucrative job offer from a large charitable organization. In her new career she meets a wonderful person, falls in love and gets married. The rest of her life turns out happy, prosperous, and fulfilling. Had she stayed put outside the burning building, she would likely have continued her below-average life in socially and economically limited circumstances.

\textsuperscript{14} Heathwood assumes that a life will, ceteris paribus, have a positive net-balance of desire satisfactions over frustrations.

\textsuperscript{15} An approach based on delineating a ‘true self’ by Conny Rosati does not seem to me to overcome the issue (it still assumes a , and a very recent proposal by Von Kriegstein appears to me to be too vague to clearly solve the problem. See Rosati (2009); Von Kriegstein (forthcoming).
Applying HDS’s method, the choice of saving the children results in an overall much higher amount of net satisfaction of desires than the alternative. What looks like a definite act of self-sacrifice was, according to HDS, in Michelle’s best interest, and thus no sacrifice after all. HDS gets it wrong.

IV.

What exactly goes wrong for HDS in Happy Ending Cases? Their distinctive feature is that there are benefits that accrue to the agent after the sacrifice, which not only outweigh the losses incurred, but also surpass the benefits of a counterfactual alternative life. It seems that our intuitive assessment of an action as self-sacrificial in these cases ignores some future benefits and harms.

A similar phenomenon exists in some of our moral evaluations. If someone hijacks the plane you’re on and threatens to kill you and other passengers, then even if you survive, write a book, and become rich and famous from it, you were still wronged by the hijacker. A rights violation is not erased by bringing about benefits later on. Analogously, in self-sacrifice, the harm incurred is not erased by later benefits. But how can we count only some of the benefits and ignore others?

It cannot simply be that we discount benefits that lie in the future. For suppose Michelle becomes famous as soon as she enters the building; and all related benefits accrue to her while she is saving the children. None of this would make us less impressed with her or alter our judgment that she is committing an act of self-sacrifice. Here we ignore even instantaneous benefits, so

---

16 This example is borrowed from Steinbock & McClamrock (1994), p. 19.
temporal proximity cannot be the criterion by which to decide which benefits to take into consideration.

Another salient feature of Happy Ending Cases is that the benefits they introduce are unexpected. Plausibly, we only consider expected benefits and harms, when we assess whether someone committed self-sacrifice. If this is right, it explains why HDS could not account for these cases. The crucial move in applying HDS is precisely to consider satisfactions and frustrations of desires the agent fails to anticipate. This includes those unexpected desire satisfactions that a happy ending would bring.

V.

In its extant form, HDS emphasizes that the agent is not anticipating later emerging desires. By giving this up, it might be able to incorporate the restriction to expected harms and benefits, while still taking future desires into account.

Since Michelle knowingly put her life at risk, why shouldn’t we allow that the good things she expects to lose include satisfactions of desires she does not yet have? In principle, she could realize that our desires change over time, and that new experiences generate new desires, some of which may get satisfied. Giving up such satisfactions of as-of-yet nonexistent desires of her future self could therefore count towards the expected harm.

Although whatever desires she now has regarding her future are outweighed by her current desire to save the children, the total amount of harms includes the loss of satisfactions of desires she does not have, since those are part of the expected harm. By giving up on the claim that the agent does not anticipate her changing or future desires, HDS can overcome the difficulties Happy
Ending Cases pose. For satisfactions resulting from the happy ending are not expected. So HDS gets to have its cake (future desires) and eat it too (excluding net-sum upending happy endings).

Still, there are case of self-sacrifice that do not involve any unexpected benefits or harms, where HDS fails, just as conventional desire-satisfaction views would. These are what I term Serial Cases. Consider:

Aisha the serial do-gooder

Aisha routinely sacrifices her own interests to help others. Her friends implore her to say “no” once in a while, to not engage in what amounts to serial self-sacrifice, since it’s making her miserable. Her response is: “So what?” Aisha doesn’t think her own well-being is very important. She has long had a strong desire to help others, which dominates much of her life. Now Peter, an acquaintance, urgently asks her for help. This will require Aisha forgoing a night’s sleep. She is already under-slept, and she has an important meeting the next day. Helping Peter would constitute self-sacrifice on her part. But can we say this according to HDS?

Aisha’s strong desire to help others outweighs her desires for sleep and for doing well the next day. There are no unexpected harms of benefits here, unlike in the Happy Ending cases. What about future desires? One might consider Aisha’s very strong urge to sleep at the time her alarm goes off in the morning a new desire – stronger than the desire she had the night before to get

---

17 I refrain from assigning a different acronym to the resulting view, because the core of HDS remains untouched. Nevertheless, the modification does diverge from points Heathwood emphasizes. I take it that the move suggested here is a change at least in spirit, if not in letter, from the original proposal.

18 If one holds a view according to which there are “standing” desires, which stay with us even when we’re not currently feeling them, this would not count as a new desire, but merely a so-called “occurrent” desire, which is a standing desire become conscious, so to speak. See e.g. Schroeder (2004). Such a view – not uncommon in the
enough sleep that night.\textsuperscript{19} This would be the kind of future desires that Aisha does not have yet but which she might foresee, so that HDS could count it among expected harms and benefits. However, given the person Aisha is, we can expect that when her alarm goes off and she acutely feels the urge for more sleep, she still stands by her decision. Her desire to help is strong enough to outweigh even those desires that arise the next day. Given Aisha’s character, it is highly likely that her desire to help will continue to dominate her motivational set. It is not impossible for her to change, but highly unlikely. Her desire to help others is a stable character trait. And there will be opportunities to satisfy it practically every day. Any given set of desires she would have in the future will be outweighed by her desire to help others.\textsuperscript{20} Aisha will continue her pattern of regular self-sacrifice. Each time, she will satisfy her strongest desire (to help others) and frustrate some weaker desires (regarding her comfort, health, etc.), resulting in a life of very high net desire satisfaction. According to HDS, her life of continuous self-sacrifice ranks very high on the scale of good lives. This result is absurd. While Aisha’s life is perhaps morally exemplary, it is not good for her. HDS gets it wrong.

\textbf{VI.}

Whether an action counts as self-sacrifice is in large part determined by our assessment of what the agent is willing to do. Michelle is willing to risk her life in order to save those kids, and Aisha is willing to help a mere acquaintance in need, accepting considerable inconveniences to herself. They are willing to commit self-sacrifice. Since self-sacrifice involves a loss in well-being literature – would cause additional difficulties for HDS. I will leave this issue aside, since I don’t want to rest my arguments on the acceptance of a particular view of desires.

\textsuperscript{19} Heathwood cites examples of this kind in his discussion of future desires, in the context of a “non-diachronic” case of self-sacrifice. See his (2011), p. 34.

\textsuperscript{20} If HDS counts some of those future desires against a current desire to help, it runs the risk of double counting. In a well-being calculus, each instance of desire frustration (or satisfaction) should only be entered once.
for the agent, a theory of well-being should be able to explain the cases of Michelle and Aisha. It should be able to say that they are willing to incur a net-loss in their own welfare. This verdict can come apart from whether their action was “in their best interest” all things considered, or in the long run.

The fact that our assessment of an agent’s loss in welfare with regard to a particular action can come apart from our all-things-considered, long-term judgment about that same agent, shows that a theory of welfare needs to have the flexibility to deliver ‘local’, as well as long-term, verdicts of people’s well-being. HDS lacks that flexibility. All it can do is pronounce an action beneficial or harmful based on the overall, long-term, well-being of an agent who does or doesn’t perform it. And standard desire theories, which do have that flexibility, cannot make sense of self-sacrifice. If there is any hope for a desire-satisfaction centered theory to be successful here, it has to be substantially different from those discussed in the literature so far. In what follows, I will propose such a different theory.

Suppose we took as our starting point not the entire, vast, web of desires each of us has, but instead focus on a small set of desires that are particularly central in those webs; some general desires that seem to be connected to most of our other, not-so-central desires. Desires such as those for love and affection, self-expression, self-determination, nutrition, warmth, bodily integrity, connection to others, safety, sex, pleasure (and no pain), exercise of our skills and talents, etc. They seem to bear some connection to all those everyday desires we usually first think of when we hear of desire-satisfaction. For example, the desires for love and affection, sex, and connection to others, all seem to be in play when we ‘select’ romantic partners. Self-determination comes in when we desire to choose our own partner, rather than to accept the one our parents suggest for us. Self-expression and self-determination play a role when children start developing preferences.
for which clothes to wear. They may be one or two dozen of these kind of central desires – let’s call them *Core Desires* – and it is plausible that most of them are shared by most humans. I propose that one’s level of well-being is determined by the balance of satisfaction over frustration of one’s Core Desires, and that all other, ‘non-core’ desires contribute to one’s well-being only insofar as their satisfaction contributes to the satisfaction of one’s Core Desires. In other words, only the satisfaction (or frustration) of Core Desires is ultimately relevant for one’s well-being.

This kind of view avoids the problem of self-sacrifice, because it ignores non-core desires, like Michelle’s desire to save the children. The fact that that is her motivationally strongest desire at the time becomes irrelevant to the question of what is good for her. What matters is the balance of Core Desire satisfaction over frustration. This balance will come out negative, since her severe and lasting injuries will frustrate many of her Core Desires. In similar fashion, the Core Desire view can deal with many other common objections to desire-satisfaction views.

Aside from its potential to overcome or avoid problems that have hampered other desire-based views, here are some independent reasons why such a view is plausible and even intuitive:

1. When giving advice about what someone should do in particular situations, or with their life in general, people often say things like: “Yes, I know you just don’t want much responsibility and an easy life, but what do you really want?”. Another common expression is: “What is it you want *deep down*?” These locutions, while perhaps cliché, and certainly vague, nevertheless point to a widely shared intuition that what people should do, if they want to do what’s good for them, is not to satisfy their momentary desires, or those desires that come most readily to mind, but instead search for some more fundamental desires that may bear some relation to those more ‘superficial’ wants, but which are somehow closer to what is in fact good for them.
2. The notion that there are some desires which are in some sense more ‘fundamental’ than others, coupled with the view that those should have some normative priority over the less fundamental ones, is held implicitly or explicitly by many philosophers. John Rawls, for example, espouses this view, when explaining his idea of a rational life plan: “[T]he rational plan for a person is the one…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.”

3. When someone close to us has a desire for X, which we don’t share, we might ask them what it is about X that they find desirable. Whether or not we ask them, most would take it to be a legitimate question, in most cases. It seems perfectly reasonable to ask this, and we may have reasonable expectations that there is in fact an answer to such a question, even if it requires prolonged reflection on the part of the addressee. Suppose the answer is “a, b, and c”, then we can infer that the person who desires X also desires a, b, and c. What’s more, if X was devoid of a, b, and c, we can assume that the person would not desire X. In other words, we commonly assume that when people have a desires for X, they also have other desires for particular things that they expect to get from X, and that ultimately it is those desires that they want satisfied.

This is of course merely the sketch of an idea in its broadest outlines. But I hope I have said enough to indicate its general plausibility and potential. I submit it as a proposal for the direction desire theorists should move from here.

---

WORKS CITED:


- (2011). “Preferentism and Self-Sacrifice,” *Pacific Philosophical Quarterly* 92, 18–38


