Preferring to Decrease One's Own Well-Being

JOHN BRONSTEEN
Loyola University Chicago

Many scholars have argued that well-being is the satisfaction of preferences, or of fully informed preferences, or of fully informed preferences about one's own life. But none of those theories can be true if it is possible to prefer, with full information, to decrease one's own well-being. And because it is possible to have such a preference, those theories cannot be true.

INTRODUCTION

Preference satisfaction is the most widely held theory of well-being.[1] Here I discuss a counterexample to that theory: a preference to decrease one's own well-being. This counterexample has been mentioned by Richard Kraut[2] and a few others, usually in passing, but it has not been used to construct a fully articulated argument against preference theories and has therefore been largely ignored by both critics and proponents of those theories. I offer such an argument, which proceeds first by explaining that it is possible to have a preference to decrease one's own well-being, and then by showing that that possibility is incompatible with three of the most prominent preference theories: the actual preference theory, the fully informed preference theory, and Derek Parfit's Success Theory.

THE STATE OF THE LITERATURE

The simplest version of the preference theory is that well-being is the satisfaction of actual preferences.[3] A far more common version is that well-being is the satisfaction of fully informed preferences.[4] A third

[1] L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford, 1996), p. 122; D. Haybron, *The Pursuit of Unhappiness* (Oxford, 2008), p. 34.

[2] R. Kraut, 'Desire and the Human Good', *Proceedings and Addresses of the American Philosophical Association* 68 (1994), pp. 39–54, at 40-1.

[3] M. Murphy, 'The Simple Desire-Fulfillment Theory', *Noûs* 33 (1999), pp. 247–72; C. Heathwood, 'The Problem of Defective Desires', *Australasian Journal of Philosophy* 83 (2005), pp. 487–504.

[4] J. Rawls, *A Theory of Justice* (Cambridge, MA, 1971), p. 408; R. Brandt, *A Theory of the Good and the Right* (Amherst, NY, 1979), ch. 6; J. Harsanyi, 'Morality and the Theory of Rational Behavior', *Utilitarianism and Beyond*, ed. A. Sen and B. Williams (Cambridge, 1982), pp. 39–62; P. Railton, 'Facts and Values', *Philosophical Topics* 14 (1986), pp. 5–31.

© Cambridge University Press 2016
Utilitas Vol. 29, No. 1, March 2017
doi:10.1017/S0953820816000212

CrossMark
https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

version, known as Parfit's Success Theory, is that well-being is the satisfaction of fully informed preferences about one's own life.[5]

All three of those versions are incompatible with the possibility that someone could have a fully informed preference about her own life that, if satisfied, does not increase her well-being. Because I will discuss an example of such a preference, I will group those three theories together and, for conciseness, refer to them collectively as 'preference theories'.

A typical objection to preference theories is that someone could have a fully informed preference about her own life that, if satisfied, does not seem *intuitively* to increase her well-being – such as Simon Keller's example of a preference for eating gravel.[6] On a few occasions, this objection has been made via examples in which even the preferror herself may not believe that getting what she wants will increase her well-being. Stephen Darwall gives an example of a woman who prefers to marry a man she does not love because she has promised to do so.[7] Matthew Adler imagines a 'penitent wrongdoer who develops a moral preference that he suffer pain and anguish as punishment for the wrongdoing'.[8] And Fred Feldman argues that someone could have a preference for displeasure.[9]

One possible strategy that a proponent of preference theories could use to cope with these counterexamples is to amend preference theories such that they conceive of well-being as the satisfaction of only those preferences (or only those fully informed preferences, or only those fully informed preferences about one's own life) that are self-interested. I will not evaluate this strategy here because my focus is on versions of preferentism that do not limit the sphere of relevant preferences to self-interested preferences. I will merely mention, though, that several leading contributors to the well-being literature have noted the obvious objection to the strategy of including only self-interested preferences in the well-being calculus, calling such a strategy 'patently circular',[10] 'viciously circular',[11] or 'in a way . . . circular'.[12]

A different strategy that a proponent of preference theories could use to cope with intuition-based counterexamples such as a preference

[5] D. Parfit, *Reasons and Persons* (Oxford, 1984), p. 494.

[6] S. Keller, 'Welfare as Success', *Noûs* 43 (2009), pp. 656–83, at 658.

[7] S. Darwall, 'Self-Interest and Self-Concern', *Social Philosophy and Policy* 14 (1997), pp. 158–78, at 164-5.

[8] M. Adler, *Well-Being and Fair Distribution* (New York, 2012), p. 180.

[9] F. Feldman, *What Is This Thing Called Happiness?* (Oxford, 2010), pp. 67–9.

[10] Sumner, *Welfare*, p. 135.

[11] R. Arneson, 'Human Flourishing Versus Desire Satisfaction', *Human Flourishing*, ed. E. Frankel Paul et al. (Cambridge, 1999), pp. 113–42, at 124.

[12] J. Griffin, *Well-Being: Its Meaning, Measurement, and Moral Importance* (Oxford, 1986), p. 22.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

for eating gravel is to bite the bullet by saying that satisfying such preferences does increase one's well-being notwithstanding any intuitions to the contrary. Because of the other appealing features of preference theories, those theories' proponents are willing to accept the counterintuitive conclusion that in these examples, one's well-being actually does increase.

Given this state of the literature, it would be important to find a counterexample to preference theories that does not rely on intuitions about whether the satisfaction of a given preference actually increases one's well-being. Perhaps surprisingly, such an example already exists: a preference to decrease one's own well-being. This example was first proposed by Kraut but has received little attention.[13] Kraut himself declined to develop an argument based on it, saying instead: 'But rather than pursue this idea, I will turn to another type of objection.' Since then, most mentions of the example have similarly been in passing.[14]

Only Ben Bradley discusses the example in detail, but his aim is primarily to show that it applies not just to preference theories but also to any theory that says one's well-being depends on the truth of one's beliefs or attitudes.[15] He also discusses whether such truth-dependent theories could avoid the example by, among other possibilities, conceiving of well-being as only those pleasures or desire satisfactions that are 'worthy'.[16] These are interesting topics, but I will not discuss them here except to note that the well-being theories that Bradley says are incompatible with the example do not include the most prominent alternatives to preference theories, which are traditional

[13] Kraut, 'Desire', pp. 40-1.

[14] See F. Feldman, *Pleasure and the Good Life* (Oxford, 2004), p. 17; R. Adams, *Finite and Infinite Goods* (New York, 1999), pp. 89–91; T. Carson, *Value and the Good Life* (Notre Dame, IN, 2000), p. 92. Feldman says only that the example seems to force a paradox upon the theory that well-being is the satisfaction of a person's actual preferences. Adams says that a preference theory cannot avoid the example by excluding from the well-being calculus people's preferences against their own well-being, because such a theory would be circular and would not sufficiently honour people's actual preferences. Carson argues that the example is not problematic for the preference-satisfaction theory of *value*, as opposed to the preference-satisfaction theory of *well-being*, because 'a bad person who punishes herself' may increase the value of her life even as she decreases her well-being.

[15] B. Bradley, 'A Paradox for Some Theories of Welfare', *Philosophical Studies* 133 (2007), pp. 45–53.

[16] In addition, Bradley supplies a response to Chris Heathwood, who argues that the paradox of wanting not to get what one wants exists regardless of whether well-being is preference satisfaction, and that therefore the example is not specifically a problem for preference theories of well-being but instead a general problem like the liar paradox. See Heathwood, 'The Problem', pp. 501-3. Bradley rebuts that argument by pointing out that only if preference theories are true does wanting to decrease one's own well-being equate with wanting not to get what one wants. Therefore, only preference theories must claim that the apparently non-paradoxical preference to decrease one's own well-being is in fact paradoxical. See Bradley, 'A Paradox', pp. 51-2.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

hedonism and typical objective-list theories. Thus those prominent alternatives are not contradicted, in the way that preference theories are, by the example of a preference to decrease one's own well-being.

My contribution here is as follows. To assess the significance of this counterexample, one must consider several questions: (1) Is it possible for someone to prefer to decrease her own well-being? (2) If so, could that preference be fully informed? (3) Is the possibility of such a fully informed preference incompatible with preference theories? I will explain why I believe each of these questions is best answered 'yes'.

MY ARGUMENT

According to preference theories, a person's well-being is increased by satisfying any of her fully informed preferences about her own life.[17] But the satisfaction of a fully informed preference to decrease one's own well-being – a preference that is obviously about one's own life – necessarily decreases one's well-being rather than increasing it. Thus, if it is possible for someone to have a fully informed preference to decrease her own well-being, then preference theories cannot be true.[18]

Is it possible for someone to have a fully informed preference to decrease her own well-being? One way to answer 'no' would be to say that well-being cannot be the object of a preference. But well-being is widely understood to be one of the most common objects of people's preferences. For Tom to want what is best for his children, for example, is for him to have a preference for their well-being. If it is possible to have that apparently normal preference, then well-being can be the object of a preference.

Indeed, most theories of well-being, including preference theories, can easily explain many preferences that have well-being as their object. If Tom wants what is best for his children, then a preference theory holds simply that Tom wants his children to get what they want. There is no problem in that. Hedonic theories and objective-list theories – the two main alternatives to preference theories in Parfit's oft-cited taxonomy[19] – also account easily for this preference. A hedonic theory of well-being holds that if Tom wants what is best for his children, then he wants his children to have pleasure. An

[17] I will discuss below the issue that arises if the satisfaction of one such preference causes other such preferences to go unsatisfied.

[18] Such a preference is incompatible with preference theories whether or not the preference is satisfied. According to preference theories, Sarah's level of well-being would be lower if any one of her fully informed preferences about her own life were unsatisfied than if that preference were instead satisfied. But if Sarah has a fully informed preference to decrease her own well-being, then her level of well-being would necessarily be *higher* if her preference were unsatisfied than if it were satisfied.

[19] Parfit, *Reasons*, p. 493.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

objective-list theory of well-being holds that if Tom wants what is best for his children, then he wants his children to have things like health, freedom and loving relationships. All of these accounts of Tom's preference are unproblematic.

Now consider the presumably common preference to increase one's own well-being: Tom wants what is best for himself. Hedonic theories of well-being say that such a preference entails wanting pleasure, objective-list theories of well-being say that such a preference entails wanting things like health, and preference theories of well-being say that such a preference entails wanting to get what one wants. There are no apparent difficulties with any of those accounts.

It thus seems possible and indeed unremarkable for someone to have a preference about well-being. But is it possible to prefer to decrease one's own well-being? On a conceptual level, the answer certainly seems to be 'yes'. If it is conceptually possible for Tom to want what is best for himself, then why would it not be conceptually possible for him to want what is worst for himself? People can want *others'* well-being to decrease: Tom might wish ill on his enemies. To be sure, preferring that one's own well-being decrease is far less common than preferring either that one's own well-being increase, that someone else's well-being increase, or that someone else's well-being decrease. But recall Adler's example of a guilt-ridden wrongdoer who wants to suffer for his misdeeds because he believes on moral or religious grounds that he deserves his life to go badly for him.[20] Or simply consider a self-loathing person. Those examples are hardly nonsensical or unimaginable. And indeed, if well-being is conceived the way that hedonic theories or objective-list theories conceive it, then there is no difficulty with a preference to decrease one's own well-being: it is clearly possible conceptually for someone to want to decrease her pleasure or worsen her health.

Hedonic theories and objective-list theories can thus make sense of preferences for or against others' well-being, as well as preferences for or against one's own well-being. Preference theories can also make sense of preferences for or against others' well-being, as well as preferences for one's own well-being, but they cannot make sense of a preference against one's own well-being. The validity of preference theories depends, therefore, on the claim that a preference against one's own well-being is impossible. And that claim, at least on a conceptual level, is false.

So a preference theorist might instead contend that although having such a preference is conceptually possible, it is nonetheless impossible as a matter of psychological fact. But such a contention would face

[20] Adler, *Well-Being*, p. 180.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

two problems. First, it too is false: the examples of penitent sinners and self-loathing people are realistic. And it seems worth noting that if the truth of preference theories depended on the truth of controversial psychological claims similar to psychological egoism, then preference theories would probably lose many of their adherents. Second, even if human psychology precluded – as a matter of contingent fact – a preference to decrease one's own well-being, the mere conceptual possibility of such a preference would still entail that well-being does not equate necessarily with the satisfaction of fully informed preferences.[21]

It thus seems possible for someone to prefer to decrease her own well-being. And if that is possible, then it is also possible for such a preference to be fully informed. The full-information requirement does not filter preferences by their content but instead functions as follows. If Jill wants to eat an apple because she does not know it is poisoned, then her preference is not fully informed because she would not have held the preference if she had known additional information. Indeed, as Wayne Sumner explains, Jill's desire to eat the apple is not accurately described as a preference at all: Jill does not actually prefer to eat the apple.[22] Thus, the full-information requirement never forecloses the possibility of a preference based on the substantive nature of the thing preferred, but instead merely insures that only true preferences are counted. Anything that can be a preference, therefore, can be a fully informed preference. So if a person could have a preference to decrease her own well-being, then she could have a fully informed preference to decrease her own well-being.

If it is possible for someone to have such a preference, then presumably a preference theorist must claim that the possibility of that preference is compatible with preference theories. I will call that claim the compatibility argument, and one version of it might go like this:

> A preference to decrease one's own well-being is a preference to have one's *other* preferences go unsatisfied. Thus, if a preference to decrease one's own well-being is satisfied, then one's well-being decreases, but only because one's preferences on balance go unsatisfied. Therefore, preference theories are compatible with a preference to decrease one's own well-being because if such a preference is satisfied, then one's well-being decreases not due to the

[21] Indeed, a claim of psychological impossibility would run counter to the way that well-being theories are typically assessed. For example, Nozick's experience-machine counterexample to hedonism does not depend on whether human beings are physically or psychologically capable of tolerating an experience machine. See R. Nozick, *Anarchy, State, and Utopia* (New York, 1974), pp. 42-5.

[22] Sumner, *Welfare*, p. 131.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

satisfaction of that preference but instead due to the non-satisfaction of other preferences.

The problem with this argument is that it breaks down if a person prefers to decrease her own well-being above any and all conflicting preferences. To see this, first consider an analogous case: Jason wants to eat thirty candy bars every day and also to lose weight. Suppose it is impossible for Jason to satisfy both wants: they are mutually exclusive. Is Jason better off (a) eating the candy bars and not losing weight or (b) losing weight and not eating the candy bars? According to preference theories, the answer depends on whether Jason prefers (a) or (b). For example, if Jason prefers (a), then his well-being is greater if he eats the candy bars than if he loses weight. The same reasoning applies to Ellie, who wants to decrease her own well-being and also wants to grow taller. Suppose it is impossible for both wants to be satisfied: Ellie can either (A) decrease her own well-being and not grow taller or (B) grow taller and not decrease her own well-being. Suppose that, as between the two possibilities, Ellie prefers (A). In that case, preference theories say that Ellie's well-being is greater if (A) she decreases her own well-being and does not grow taller than if (B) she grows taller and does not decrease her own well-being. But this brings us back to the original problem: Ellie's well-being cannot be greater if it decreases than if it does not decrease.[23]

The compatibility argument would not be improved by differentiating first-order preferences from second-order preferences, where the latter are defined as preferences about one's preferences. A preference theory might make use of this differentiation by claiming, for example, that well-being is the satisfaction of first-order preferences. According to such a theory, a preference to decrease one's own well-being would be a second-order preference – a preference not to satisfy one's first-order preferences. And thus, the argument would go, it is unproblematic that the satisfaction of such a preference decreases one's well-being. If Tim prefers to decrease his own well-being, and if that preference is satisfied, then Tim's well-being is said to decrease not due to the satisfaction of Tim's preference to decrease his own well-being (a preference that is, according to this theory, irrelevant to the calculus of Tim's well-being because it is a second-order preference) but instead due to the non-satisfaction of Tim's first-order preferences. Thus, such a version of preferentism seems compatible with the possibility of a preference to decrease one's own well-being. But there are three

[23] Whatever other wants Ellie may have, we may hypothesize that she prefers to decrease her own well-being above all of those other wants combined. However unusual that might be, it is surely possible.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

problems with that argument. First, if Tim's *only* preference were to decrease his own well-being, then satisfying that preference would necessarily decrease his well-being, and the decrease could not be attributed to the non-satisfaction of his other preferences (his first-order preferences) because he has no other preferences. This means that Tim's well-being would decrease for a reason other than the non-satisfaction of any of his preferences, and that is inconsistent with preference theories. Thus, even if a preference theory excludes second-order preferences from the well-being calculus, the theory must still deny that it is possible to prefer only to decrease one's own well-being. And that denial is problematic for the reasons given above.[24] Second, even if a preference to decrease his own well-being could not be Tim's only preference, it could surely be his strongest preference. And to treat the satisfaction of Tim's strongest preference as irrelevant to Tim's well-being is inconsistent with conceiving of well-being as preference satisfaction. Put more generally, something is no less a preference by virtue of its being a second-order preference.[25] Third, typical second-order preferences are preferences to change one's first-order preferences – such as a preference to stop preferring to smoke cigarettes.[26] And the satisfaction of such second-order preferences clearly affects one's well-being, so excluding those preferences from the well-being calculus would make a theory yield indefensible results.

A different version of the compatibility argument has been advanced by Brad Skow, who says that if a preference can be partially satisfied, then preference theories are compatible with a preference for one's life to go badly.[27] (I will assume for the sake of argument that a preference for one's life to go badly is equivalent in relevant respects to a preference to decrease one's own well-being. Only if that is true would Skow's claim be a rejoinder to the argument I have laid out.) Skow's argument begins by observing that preference theories make this claim:

> *P*: A person's level of well-being is the same as her level of preference satisfaction.[28]

[24] For example, only preference theories, and not hedonic or objective theories, must deny that it is possible to prefer only to decrease one's own well-being. And a claim of conceptual impossibility seems insupportable, whereas a claim of psychological impossibility seems insufficient.

[25] Indeed, Peter Railton privileges second-order preferences because doing so insulates preference theories from various criticisms. See Railton, 'Facts'.

[26] See Bradley, 'A Paradox', p. 50.

[27] B. Skow, 'Preferentism and the Paradox of Desire', *Journal of Ethics and Social Philosophy* 3 (2009), pp. 1–16.

[28] As I do throughout, I am referring only to fully informed preferences about one's own life. And in my discussion of Skow, I am also following Skow's useful simplifying assumption that the preference for one's life to go badly is one's only preference. The

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

If Margot's only preference is to eat a pie, then she has zero well-being if she eats none of the pie, total well-being (call it '100') if she eats all of the pie, and (according to Skow) half-well-being ('50') if she eats half of the pie. Now suppose Margot's only preference is instead for her life to go badly. If she *half*-satisfies that preference, then her level of well-being is 50, which is the same as her level of preference satisfaction, thereby satisfying *P*. Skow says this shows that preference theories are compatible with the possibility of a preference for one's life to go badly.

But what if Margot prefers only for her life to go badly and satisfies that preference to any degree *other* than 50 per cent? If she satisfies that preference 100 per cent, then her well-being is 0; if she satisfies it 0 per cent, then her well-being is 100; if she satisfies it 30 per cent, then her well-being is 70; and so on. In any case other than a level of 50 per cent satisfaction, Margot's level of preference satisfaction differs from her level of well-being, and that violates *P*. So if a preference theory relies on Skow's idea to accommodate the example of a preference for one's life to go badly, then that theory must claim it is impossible for someone to satisfy such a preference to any degree other than 50 per cent. And that claim is no more plausible than the claim, required of non-Skow preference theories, that it is impossible for someone to prefer for her life to go badly.[29] Presumably, the only argument for believing either of those claims is that the truth of preference theories depends on them. But that argument is tautological because the matter in question is whether preference theories are true.

argument would be unaffected if one had other preferences as well, as long as one preferred for her life to go badly more strongly than she preferred those other things collectively.

[29] As discussed above, an advantage of hedonic and objective theories is that they need not insist that a preference for one's life to go badly differs radically from other seemingly similar preferences such as a preference for one's life to go well or a preference for others' lives to go badly or to go well. That advantage is undiminished if preference theories shift from claiming that it is impossible to prefer one's life to go badly to claiming that it is impossible to satisfy such a preference to any degree other than 50 per cent. According to hedonic and objective theories, it is possible to satisfy a preference for or against one's own well-being, or a preference for or against others' well-being, to any degree.

Moreover, it seems highly unusual for *any* preference to be satisfiable only to a degree of 50 per cent. To be sure, some preferences may be satisfiable only to a particular degree due to physical laws or other restrictive circumstances. For example, if Martin prefers to live on the sun, then his preference can never be satisfied to any degree other than zero. But examples like Martin's preference lend no support to the claim that a preference for one's own life to go badly can only be precisely half-satisfied, because the reason supporting that claim (unlike the reason – physical laws – in examples like Martin's preference) is merely the asserted truth of preference theories themselves. And the assertion that preference theories are true cannot, without circularity, be used to substantiate the truth of preference theories.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

A preferentist could instead advance yet another version of the compatibility argument, relying on an idea of Phillip Bricker's.[30] Bricker proposed that someone's lifetime well-being is the level of satisfaction of the preferences held by her past, present and future selves.[31] For example, suppose Gail spends her first 40 years of life preferring that she will *never* work for a big corporation. Then, at the age of 40, Gail changes her mind and does prefer to take such a job. According to Bricker, the satisfaction of Gail's preference at the age of 40 actually decreases Gail's lifetime well-being because it reduces Gail's overall lifetime level of preference satisfaction. This is because it causes Gail's previous preferences against corporate work to go unsatisfied, and she held those preferences much longer than – and, let us assume, just as intensely as – her contrary preference at the age of 40. (Let us also assume that Gail was going to die tragically at the age of 41 whether or not she took the job, so we need not consider her future selves' preferences.) Now consider Ryan, who prefers at the age of 40 to decrease his own well-being. Like Gail, Ryan may have had preferences at other times of his life that conflicted with his preference at the age of 40. And it is also possible that the satisfaction of Ryan's preference at the age of 40 would change Ryan's future in a way that affects his lifetime quantum of preference satisfaction. For these reasons, it is possible that satisfying Ryan's preference to decrease his own well-being could reduce rather than increase his lifetime level of preference satisfaction, thereby making Ryan's reduction in well-being (which is tied inextricably to satisfying the preference) compatible with preference theories. By linking well-being to the satisfaction of all preferences rather than to the satisfaction of any individual preference, Bricker's approach frees preference theories from being forced to claim that satisfying a given preference must increase the preferror's well-being. And that makes preference theories, according to this argument, compatible with the example of a preference to decrease one's own well-being.

But there is a problem with that argument. To explain the problem, let me first avoid confusion by noting that Bricker's approach concerns *lifetime* well-being rather than well-being at a given time. Therefore, the most direct challenge to Bricker would be a preference to decrease one's own lifetime well-being. Let us say that Ryan holds such a preference. Bricker's preferentism can accommodate that case if, by

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

[31] Bricker's calculus also includes alternative versions of oneself that would have existed if one had made different choices. Thus, satisfying a current preference could diminish one's lifetime well-being if such satisfaction leads to a future with less preference satisfaction than one would otherwise have had.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

satisfying the preference, Ryan decreases his lifetime level of preference satisfaction. That happens if most of Ryan's selves prefer not to decrease his lifetime well-being, in the same way that most of Gail's selves preferred not to work for a big corporation. But the problem is this: although it is *possible* that most of Ryan's selves prefer not to decrease his own lifetime well-being, that is not *necessarily* the case. Just as a standard preference theory must claim that it is impossible for Ryan to prefer to decrease his own well-being, Bricker's theory must claim that it is impossible for most of Ryan's selves to prefer to decrease Ryan's own well-being; and the latter claim is no more plausible than the former one. Suppose Ryan commits a grievous sin at the age of 40. Believing it is right for himself to suffer for the sin, the 40-year-old Ryan prefers that his lifetime well-being decrease from its current level. And suppose, as we did with Gail, that Ryan will die fairly soon (just after the preference gets satisfied or not) of an unrelated cause regardless of whether this preference is satisfied.[32] It could easily be the case that most of Ryan's younger selves shared the preference of his 40-year-old self: they too preferred for Ryan to suffer if he were ever to commit a grievous sin. And if that is the case, then satisfying Ryan's preference to decrease his own well-being *increases* Ryan's lifetime level of preference satisfaction while necessarily decreasing his lifetime level of well-being. Thus the incompatibility between preference theories and this counterexample is not solved by Bricker's version of preferentism.

So it is possible to have a preference to decrease one's own well-being, and that possibility is incompatible with preference theories. Could preference theories simply make an exception for this one preference? For example, the simple preference theory might become: 'Well-being is the satisfaction of preferences, except for a preference to decrease one's own well-being.' The problem with such an approach is, in my view, that it is analogous to that of a hedonist who responds to Nozick's experience-machine counterexample by amending the hedonic theory to say this: 'Well-being is happiness, except if you're on an experience machine.'[33] To a large degree, the core idea of preference theories is that getting what you prefer makes you better off, no matter what you prefer.

[32] This makes the case simpler but is not required for the argument to work. There is no reason to believe that if Ryan lives far beyond the age of 40, satisfying his preference to decrease his own well-being will necessarily reduce his level of preference satisfaction at future ages or in alternative versions of himself.

[33] The example of a preference to decrease one's own well-being actually poses a more intractable problem for preference theories than does the experience machine for hedonic theories. Hedonists can and often do bite the bullet by claiming that a life on the experience machine would be high in well-being. But when a preferentist faces the counterexample discussed here, he cannot bite the bullet because satisfying a preference to decrease one's own well-being necessarily does not increase one's well-being.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

The fact of preferring it, not the nature of the thing preferred, is what matters. (Hedonism analogously claims that becoming happier makes you better off, no matter what makes you happier.) Preference theories, like hedonic theories, thus say not just what increases well-being but why those things increase well-being. And preference theories' answer is that *X* increases your well-being *because* you prefer *X*. So if something does not increase your well-being even though you prefer it, then that undercuts the core idea of preference theories in a way that can't be solved by excluding such a preference by fiat. At the very least, a preferentist would need to provide a principled reason for making the exception,[34] and it is not clear what that reason could be.

In addition, a preference to decrease one's own well-being may not be as obscure, or as easy to dismiss as unique, as it might seem. For example, Daisy might prefer that her child prosper through Daisy's self-sacrifice.[35] Preferences of that form seem not only possible but perhaps even somewhat common, and they are incompatible with preference theories in the same way as is a simple preference to decrease one's own well-being. Excluding all forms of this species of preference from the well-being calculus thus constitutes more than an innocuous tweak to a preference theory. It is a substantive change that requires justification.

Preferring that one's child prosper through one's self-sacrifice is an example of the species of preference I have been discussing in this article. A different species of preference is 'ordinary' self-sacrificial preferences, which simply involve prioritizing something over one's own well-being. There is a small literature on whether such self-sacrifice poses a problem for preference theories. Mark Overvold is typically credited with raising this issue, and a few others including Stephen Darwall have fleshed it out.[36] The claim is that if Jenny's well-being were the extent to which she gets what she prefers, then it would be impossible for Jenny to prefer to sacrifice her own well-being for some other value such as, say, the well-being of her loved ones. And because it is possible for Jenny to prefer such self-sacrifice, preference theories cannot be true. Whether that argument succeeds is disputed because, as noted above, it can be met with a rejoinder that bites the bullet.

[34] Even Parfit's Success Theory must do so to explain why it excludes preferences that are not about one's own life. And although it might be easier to find a principled reason for ruling out such preferences than for ruling out a preference to decrease one's own well-being, this burden still weighs heavily on the Success Theory and may be a reason that it has not been embraced nearly as widely as the full-information preference theory. The full-information preference theory is not similarly undercut by excluding uninformed preferences because, as explained above, those can plausibly be claimed not to be actual preferences.

[35] I thank an anonymous reviewer for this point.

[36] M. Overvold, 'Self-Interest and the Concept of Self-Sacrifice', *Canadian Journal of Philosophy* 10 (1980), pp. 105–18; Darwall, 'Self-Interest'.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press

For example, Chris Heathwood deploys such a rejoinder by imagining a woman named Alice, who faces a choice on a Friday night between working in a soup kitchen and going dancing.[37] Heathwood says that the universe of Alice's relevant preferences includes both what she would have preferred if she had gone dancing and what she would have preferred if she had worked in the soup kitchen; and that if, considering all of those preferences, Alice receives more preference satisfaction from working in the soup kitchen than she would have received from going dancing, then 'it is no longer very intuitive that Alice is doing what is worse for her on this Friday night'. The point I would like to make is that, whether or not Heathwood succeeds in defending preference theories from the problem of self-sacrifice, his method of argument cannot be deployed against the example of a preference to decrease one's own well-being. For if Bertha prefers to decrease her own well-being, and if she receives more preference satisfaction from decreasing her own well-being than she would have received from not decreasing her own well-being, then it remains not only intuitive but also necessarily true that Bertha's well-being decreases from receiving more rather than less preference satisfaction. And that contradicts preference theories.

CONCLUSION

According to preference theories, satisfying a fully informed preference about one's own life increases one's well-being. But satisfying a fully informed preference to decrease one's own well-being cannot increase one's well-being. Therefore, preference theories are false if it is possible to prefer to decrease one's own well-being. There does not seem to be any reason to believe that such a preference would be impossible, so the best conclusion to draw is that preference theories are false.[38]

jbronst@luc.edu

[37] C. Heathwood, 'Preferentism and Self-Sacrifice', *Pacific Philosophical Quarterly* 92 (2011), pp. 18–38, at 32-5.

[38] I am grateful to Benjamin Callard, Richard Kraut, Douglas MacKay, Sonja Starr and the reviewer for their helpful comments on earlier drafts. In particular, I thank Josh Sheptow, who commented indefatigably on many drafts and without whose help this article could not have been written.

https://doi.org/10.1017/S0953820816000212 Published online by Cambridge University Press