Preferences, Welfare, and the Status-Quo Bias[1]

This is a preprint of an article whose final and definitive form will be published in the *Australasian Journal of Philosophy* 2010; the *Australasian Journal of Philosophy* is available online at: http://www.tandf.co.uk/journals/.

Dale Dorsey

Department of Philosophy
University of Kansas
Wescoe Hall, rm. 3090
1445 Jayhawk Boulevard
Lawrence, KS 66045
ddorsey@ku.edu

Preferences play a role in well-being that is difficult to escape. Though it is possible to refuse to grant ultimate authority over *a*'s well-being to *a*'s preferences, that preferences have at least some authority is strongly intuitive. But whatever authority one grants to preferences, their notorious malleability seems to cause problems for any theory of well-being that employs them. Most importantly, preferences appear to display a *status-quo bias*. Whether as a result of rational deliberation or various other psychological mechanisms, persons will come to prefer what they are likely rather than unlikely to get. Some theorists have attempted to solve these problems by constraining or limiting the relevance of certain preferences to well-being. But, as of yet, there is very little consensus that any such attempts have succeeded.[2]

In this paper, I seek to accomplish two major tasks. The first is to provide a more precise characterization of the status-quo bias and how it infects most commonly accepted theories of well-being. Second, I hope to provide an alternative characterization of an agent's preferences

[1] Doug Portmore's helpful influence on this paper is extremely difficult to overstate. I also would like to thank Nicole Hassoun, Ramona Ilea, S. Matthew Liao, and anonymous reviewers for helpful discussion and criticism. This paper also benefited from in-depth discussion in seminars at the University of Alberta and the University of Kansas.

[2] See Jennifer Hawkins, "Well-Being, Autonomy, and the Horizon Problem" in *Utilitas* 20 (2008); M. Rickard, "Sour Grapes, Rational Preferences, and Objective Consequentialism" in *Philosophical Studies* 80 (1995); Mozaffar Qizilbash, "Well-Being, Adaptation, and Human Limitations" in *Preferences and Well-Being*, ed., Olsaretti (Cambridge: Cambridge University Press, 2006).

that succeeds in avoiding the status-quo bias, a characterization I refer to as "preference coherentism". I also note what I do not seek to accomplish. Some theories will treat an agent's preferences as the ultimate arbiters of well-being. Others will grant preferences more limited authority, perhaps constrained by an objective conception of well-being or human flourishing.[3] Whether preference coherentism is plausible as a full-fledged theory of well-being is not my concern. I seek only to defend preference coherentism as a theory of *welfare-relevant* preferences—a theory of that which determines well-being whenever preferences have all-things-considered authority over prudential value.

1. Preferences, Authority, and the Status-Quo Bias

Before I begin, I wish to outline some conceptual groundwork. Call *a*'s preference for *x* over *y* at time *t* "authoritative" if and only if *a*'s preference for *x* over *y* at *t* is a sufficient condition for *x*'s being all-things-considered better for *a* than *y* at *t*. Though there are other ways of understanding the prudential authority of preferences, which I will not discuss here, this account of the authority of preferences is reflected in the views of many—mostly desiderative—theorists, including Hobbes,[4] Sidgwick,[5] Railton,[6] and many others.[7] One further point is worth

[3] Richard Kraut, in "Desire and Human Good" (*Proceedings and Addresses of the American Philosophical Association* 68 (1994)), argues that human welfare is constituted by the fulfillment of preferences or desires for things that are worth wanting in themselves. In effect, Kraut places a constraint on objective welfare goods: that they be preferred. Richard Arneson, in "Human Flourishing versus Desire Satisfaction" (*Social Philosophy and Policy* 16 (1999)), argues that there should be no such constraint on objective goods, but also argues that the fulfillment of one's desires or preferences can be one element of an "objective list" theory of human well-being.

[4] "Whatsoever is the object of any mans Appetite or Desire; that is it, which he for his part calleth Good: And the object of his Hate, and Aversion, Evill," *Leviathan*, I 6. See also *Human Nature*, VIII 3.

[5] "[I]t would have to be said that 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 forseen and adequately realised in imagination at the present point of time," *The Methods of Ethics* (Indianapolis, IN: Hackett Publishing Company, 7th ed., 1981 [1907]), 111-12.

[6] "An individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality," "Facts and Values" in *Facts, Values, and Norms* (Cambridge: Cambridge University Press, 2003), 54.

[7] See, for instance, David Lewis, "Dispositional Theories of Value" in *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000), 71; R. B. Perry, *The General Theory of Value* (Cambridge, MA: Harvard University Press, 1954), 115-6. It is worth noting that Perry's view is not specifically a theory of the good for persons, but is rather a theory of the good more generally. However,

noting. I confine myself here to discussion of—as the above definition makes explicit—the time-relative authority of preference: what is good for *a* at *t*, rather than the vexed topic of how the authority of preference issues in judgments about what is good for a person *simpliciter*, or over the course of a whole life. Hence in discussing the authority of preference, I will assume that the authority of preference holds that all potential goods will be good *at time* t.[8]

Though we might reject the view that preferences are always authoritative over well-being, it is plausible to say that *at least within certain constraints* (set by whatever preference-independent facts about welfare there may be), preferences are authoritative. The attraction of this view is easy to see. Many have claimed—plausibly—that an agent should be *sovereign* over her own good.[9] This idea has been interpreted in many ways, but one popular interpretation holds that an agent's good is at least in part determined by that agent's values.[10] But assuming that an agent's preferences constitute her values, it would appear that satisfying the demand for agential sovereignty requires us to hold that the agent's preferences are authoritative (at least when agential sovereignty is appropriate). This rationale is stated loosely, but the plausible link between *a*'s values and *a*'s well-being appears to explain the attraction of the view that, so long

---

the structure of his view betrays a status-quo bias, and hence Perry's view will be subject to the problem I explore here.

[8] To record my own view, it seems quite plausible to say that what is good for someone at *t* forms a crucial building-block for that which is good for someone over the course of an entire life. Plausibly, the quality of a person's entire life is given by the aggregate well-being score of the individual times of their lives. I shall not argue for this view; indeed it has been the subject of some criticism by, for instance, David Velleman ("Well-Being and Time" in *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000). In any event, any plausible view of well-being over the course of a life will treat a given individual's well-being at particular *times* as an important building-block, even if such a view does not adopt a simple aggregative stance. One further point is worth noting. That lifetime well-being is calculated in a way that treats a person's times of their lives impartially does not entail that prudentially rational *choice* should do so, especially in the face of the phenomenon of preference change. See, for instance, Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), ch. 8, Krister Bykvist, "Prudence for Changing Selves" in *Utilitas* 18 (2006), Dennis McKerlie, "Rational Choice, Changes in Values over Time, and Well-Being" in *Utilitas* 19 (2007), David Brink, "Prudence and Authenticity: Intrapersonal Conflicts of Value" in *The Philosophical Review* 112 (2003). In this paper, I seek only to discuss that which is in a person's well-being at *t*, rather than that which one should choose to do at *t*, or at *t*-1, given the phenomenon of changing preferences. Indeed, as I shall argue, cases of genuinely status-quo biased preferences do not present cases of *welfare-relevant* preference change at all. See §§2, 4.3.

[9] Cf. Arneson, 116.

[10] Cf. Railton, 47. See also L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Oxford University Press, 1997), ch. 2, esp. 38.

as *x* is preferred to *y* by *a* at *t*, *x* is better than *y* for *a* at *t*.

Views that ascribe authority to preferences, however, are subject to problematic biases. Consider, for instance, the problem of *sour grapes*.[11] Sour grapes cases often take the following form. Consider "The Hedonist":

> *The Hedonist*: Happy, for many years of his life (including time *t*-1), preferred great achievement in hockey to other possible activities, including hedonic satisfaction. Happy regarded hockey as a worthy pursuit in itself, and hence preferred great achievement in hockey for its own sake. However, at *t*, Happy switched his preference for hockey as a direct result of years of rejection and failure (i.e., were great achievement in hockey available, Happy would not have come to prefer the life of hedonic satisfaction). Happy now prefers pleasurable experiences to achievements in hockey.[12]

Happy's case is an example of the phenomenon of sour grapes. Like the fox in the classic fable, Happy comes to prefer the available to the unavailable. If Happy's preferences are authoritative, this would entail that hedonic satisfaction is better for Happy at *t* than great achievement in hockey. But our intuitive reactions disagree. Because Happy's preferences at *t* are a result only of an inability to be a great hockey player, it seems wrong to say that Happy's preferences should be authoritative over that which is good for him. If so, Happy's preferences appear unable to plausibly determine the relative goodness of hockey versus hedonic satisfaction in a true account of Happy's well-being.[13]

---

[11] John Elster, "Sour Grapes: Utilitarianism and the Genesis of Wants" in *Utilitarianism and Beyond*, ed. Sen and Williams (Cambridge: Cambridge University Press, 1982), and *Sour Grapes: Studies in the Subversion of Rationality* (Cambridge: Cambridge University Press, 1983), esp. ch. 3. Closely related is the phenomenon of "adaptive preferences," discussed in Martha Nussbaum, "Women and Cultural Universals" and "American Women: Preferences, Feminism, and Democracy" in *Sex and Social Justice* (Oxford: Oxford University Press, 1999), and *Women and Human Development: The Capabilities Approach* (Cambridge: Cambridge University Press, 2000), esp. ch. 2; L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Oxford University Press, 1996), esp. ch. 6; Amartya Sen, *Inequality Reexamined* (Cambridge, MA: Harvard University Press, 1992), 6-7, and several other citations by Sen.

[12] Doug Portmore suggested a helpful revision of this case.

[13] That Happy's preference is not authoritative is not evidence that achieving hedonic satisfaction is not good for him. Indeed, Happy may have some *other* preference, viz., the preference for hedons rather than *nothing*, which *is* authoritative. Furthermore, one might also claim that hedonic satisfaction is good for the person who achieves it no matter what they prefer. To say that a particular preference is not authoritative is simply to say that this preference does not establish betterness of the preferred object to the dispreferred object. It does not entail that the dispreferred object is of no welfare value. Thanks to an anonymous reviewer for bringing this point to my attention.

Happy's sour grapes preferences display what I call a "status-quo bias". A status-quo biased preference is a preference for that which will, or is likely, to occur that arises merely because its object will, or is likely to, occur. Happy prefers the hedonic life *only* because he cannot achieve the life of hockey achievement. However, it is strongly intuitive to say that, at least in this case, great hockey achievement would be better for Happy than hedonic satisfaction. We are inclined to say that, at *t*, hockey achievement is ranked as more valuable for Happy than hedonic satisfaction.

The status-quo bias is evident in classic examples of adaptive preferences. Consider an oft-quoted passage from Sen:

> A person who has had a life of misfortune, with very little opportunities, and rather little hope, may be more easily reconciled to deprivations than others reared in more fortunate and affluent circumstances. The metric of happiness may, therefore, distort the extent of deprivation, in a specific and biased way. The hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie may all take pleasures in small mercies, and manage to suppress intense suffering for the necessity of continuing survival, but it would be ethically deeply mistaken to attach a correspondingly small value to the loss of their well-being because of this survival strategy.[14]

Sen speaks of adaptive *pleasure*, but one could also imagine that any of these characters not only takes pleasure in small mercies, but in fact comes to *prefer* for its own sake the life he or she has to alternatives with greater capabilities or achievements. As Sen's case illustrates, a status-quo bias can arise in many different ways. Biased preferences can be the product of, as Sen notes, a "survival strategy". In addition, Jennifer Hawkins suggests that a lack of "self-worth" can lead an agent to prefer the shabby conditions in which they live.[15] However they arise, our intuitive judgment appears to be similar: it seems wrong to say that the hedonic life would be better for Happy, that the life of landless labor would be better for the landless laborer. Hence status-quo biased preferences appear to lack authority over an agent's welfare.

---

[14] Amartya Sen, *On Ethics and Economics* (Oxford: Blackwell, 1987), 45-6.
[15] See Hawkins, op. cit.

The status-quo bias should be distinguished from another way in which preferences can fail to be authoritative. Preferences can lack authority if one *prefers the worse*. One prefers the worse if one prefers something objectively better for its own sake to something objectively worse for its own sake. The landless laborer, quite plausibly, prefers the worse insofar as he prefers landless labor to a life of greater capabilities and achievements. But the status-quo bias is, and should be kept, distinct from the problem of preferring the worse. For instance, I might prefer landless labor, not because of a survival strategy, or because I have been conditioned to accept it as the only life I could lead, but simply because I prefer it to a life of greater opportunity, achievement, or pleasure. (Perhaps these alternative lives are even available to me.) I might simply desire to live a life that, as it turns out, is objectively bad.[16] Here I prefer the worse, but I do not display a status-quo bias. Furthermore, my preferences can display a status-quo bias without my preferring something objectively worse. I consider Happy to be a prime example; it strikes me as implausible to say that hockey achievement is any objectively better than hedonic satisfaction.[17] But to make this claim more perspicuous, consider:

> *The Painter*: Erin, at *t*-1, intrinsically preferred a life of excellent achievement in dancing to a life of excellent achievement in painting and dedicates years to the achievement of her goal. At *t*-1, Erin regarded the movement of the human form as a much more meaningful expression of her artistic temperament than mere paint on canvas. However, as a result of years of rejection and failure, she came to a decision. To avoid continued frustration and regret, she would attempt to alter her preferences away from dancing and toward painting. Suppose she is successful: Erin, at time *t*, prefers the life of painting to the life of dance.

Erin's preferences display a status-quo bias. Were she able to achieve her *t*-1 preference, she would not have come to prefer painting to dancing. Her preference is an example only of a successful attempt to avoid the predictable and painful frustration that comes along with a preference for that which is simply unavailable. It thus seems quite right to say that the life of a

---

[16] For an important example of this, see Rawls's "grass-counter" case in John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), 432.

[17] Some may deny this. See, for instance, Thomas Hurka in *Perfectionism* (Oxford: Oxford University Press, 1993), esp. ch. 3. For a critique of this view, see Dale Dorsey, "Three Arguments for Perfectionism", forthcoming in *Noûs*.

dancer would be better for Erin at *t* than the life of a painter. But her biased preferences are not an example of preferring the objectively worse to the objectively better.

Distinguishing these dual problems sheds important light on previous attempts to address the status-quo bias in the form of adaptation and sour grapes. One natural thought is that the status-quo bias can be solved by making use of an objective conception of well-being, one that imports values that are independent of preference into the prudential evaluation of lives.[18] This would solve the case of the landless laborer, for instance, by simply declaring the landless laborer's life bad in a way that is independent of preference.[19]

This proposal cannot defeat the status-quo bias because it conflates the status-quo bias and preference for the worse. This conflation leads to a failure to address cases of the status-quo bias that do not *also* involve cases of preferring the worse. *The Painter* is one such case. It is wildly implausible, it seems to me, to declare a life of excellent achievement in painting as objectively worse than a life of excellent achievement in dancing. But unless the objective view in question is willing to take the implausible step of declaring that neither possibility is better for Erin, it must treat Erin's preferences with some authority. But it is also plausible to insist that Erin's life would go better were she able to make it as a dancer than it would were she to obtain great achievement in painting. Her preference for painting is simply a result of failure: adaptation to the life that is available to her. Hence any objective view that takes the plausible step of allowing the occasional authority of preference (when, say, nothing else of weighty objective value is at stake) remains vulnerable to the status-quo bias despite its solution to the problem of preferring the worse.

---

[18] Hawkins proposes a solution that accepts that certain lives can be *bad* for persons regardless of their own preferences, but cannot be *good* for persons without the authority of their preferences. Her proposal accepts an objective solution in the sense I mean here, because she judges the objective badness of lives form a standpoint external to the agent's own values. See Hawkins, 167-8. I refrain from identifying all the different "objective" theories of value, but paradigmatic examples are defended in two different forms by Richard Arneson (the "objective list" view) and David Brink (perfectionism), in "Human Flourishing and Desire Satisfaction", op. cit., and "The Significance of Desire" in *Oxford Studies in Metaethics*, v. 3, ed. Shafer-Landau (Oxford: Oxford University Press, 2008), respectively.

[19] See, Rickard, op. cit.; Qizilbash, 95-101.

Though some biased preferences can lead one to prefer the worse to the better, the *distinctive* problem illustrated by the status-quo bias is the violation of the following plausible principle:

> *Independence*: Whether *x* would be better for *a* than *y* at *t* is independent of whether *x* rather than *y* is more likely to occur.

In principle, it ought to be possible to rank-order the prudential benefit at *t* of any possible prudential goods for any individual. The mere fact that some of these are unavailable, or are unlikely to be available, should not influence a proper evaluation of such goods. The mere fact that Happy *can* live the hedonic life should not by itself license the conclusion that hedonic satisfaction is better for Happy than hockey achievement at *t*. But because Happy's preference for the hedonic life arises because hockey playing is not available, treating Happy's preference for hedonic satisfaction as authoritative automatically imports facts of availability into the evaluation of prudential goods: hedonic satisfaction is declared better for Happy than hockey achievement.

Independence is plausible, and the status-quo bias is problematic because it violates *Independence*. Given the tendency of agents to shift their preferences toward that which they could actually achieve, the available appears to get "extra points" merely for being available. But this seems wrong. In order to give an honest account of the welfare value of potential goods, accounts of welfare should not treat availability or its lack as a *de facto* feature of prudential value. But this injunction is violated by any theory that grants even limited authority to preferences: availability *is* a factor in determining what people prefer, and hence availability, on views that allow at least some authority of preference over well-being, is a *de facto* feature of prudential evaluation.

2. Two Objections[20]

[20] This section exists thanks to the helpful and challenging questions put to my account by an anonymous reviewer.

It is worth pausing to consider in greater detail how Happy and Erin's preferences cause problems for the, at least occasional, authority of preferences over well-being. My discussion here will take the form of responses to two objections. First, consider two possible ways their *t*-1 preferences might be further specified. It might be that, for instance, Happy's *t*-1 preferences are to live a life of great hockey achievement *no matter what he actually prefers at t*. If specified in this way, we appear to be able to claim that Happy's life is worse—at least in one respect—for satisfying his sour grapes preferences: he fails to achieve that which he wanted at *t*-1, hence Happy's overall lifetime preference satisfaction is lower than it would have been had he (a) been a successful hockey player and (b) not changed his preferences. In other words, even if he achieves plenty of hedonic satisfaction at *t*, his life on the whole goes worse than it might have gone: his *t*-1 preference goes unsatisfied.

It also might be that Happy's pre-sour grapes preference is simply to live a life of great hockey achievement *only if he also wants great hockey achievement at* t (in other words, Happy's preference is "conditional on its own persistence"[21]). On the second possible interpretation, does it really seems so problematic to say that, at *t*, hedonic satisfaction is better for Happy than achievement in hockey? After all, at *t*-1, he only wants to be a hockey player so long as this preference remains.

Leave the second possibility aside for the moment. Let's simply assume Happy's preference is of the former character. This objection notes something quite true: Happy's lifetime well-being score is, on the whole, better if he achieves hockey and continues to prefer it. But this does nothing to vindicate the authority of Happy's preference at *t*. If his preference for hedonic satisfaction rather than hockey achievement is authoritative, it remains true that pleasurable experiences are better for Happy than hockey achievement *at t*. But it is precisely *this* verdict that seems so implausible: merely because Happy cannot achieve success at hockey, his preferences

[21] Cf. Parfit, 151.

adapt to that which is available, and hence this adaptation, when it comes to hedonic satisfaction versus hockey achievement, entirely shifts what would be good for him at *t*. In this way, that which would be good for Happy at *t* (assuming the authority of his sour grapes preferences) violates *Independence*.

The second objection picks up here. Why, the objection goes, should we believe that either person's welfare at *t* is determined, not by his or her preferences at *t*, but rather by his or her preferences at *t*-1? In Erin's case, it seems quite plausible to say that dancing is better than painting at *t*-1, prior to her shift. But at *t*, she shifts her goals. Why, then, shouldn't we believe that Erin's welfare at *t* is determined by her *t* preferences, rather than *t*-1 preferences, and hence declare that painting is better for Erin at *t*? This objection can be stated in slightly different terms: on the one hand, it seems right to say that Erin's preferences ought to have authority over her well-being, especially given that nothing else of objective value is in play. But on the other hand, we seek to reject the authority of her *t* preferences to determine her well-being at *t*, given their status-quo bias. We seem to want Erin's preferences to determine her welfare, and then refuse to grant her preferences authority. What gives?

In responding to this objection, recall the motivation for accepting the authority of preference. At least some of the time, we appear committed to the claim that what is good for a person should be determined by that which they *value*. But in the case of Erin, it seems right to say that her preference at *t* for painting over dancing does not really capture her evaluative perspective: her preferences are merely a result of an unusually successful attempt at strategic preference engineering. Though this intuition is hard to state with any precision, it is plausible to believe that Erin's biased preferences are not what we really care about when we think Erin's preferences should have authority over her good.

The reason Erin and Happy's cases are so problematic for the authority of preference is that these cases clearly illustrate that any particular preference need not always express its

possessor's values. Hence though it is correct to say that at least some of Erin's preferences shift from *t*-1 to *t*, this does not entail that *that which we care about when it comes to agential sovereignty over the good* also shifts. Though she develops new preferences, these preferences do not express her genuine values. Rather, they are a mere product of circumstance; they reflect, as it were, a strategic "coping mechanism". Hence even if Happy's *t*-1 preference is "conditional on its own persistence," we might still say that Happy's *t* preference lacks authority. Happy's *t* preference leaves it an open question whether his genuine values persist, or do not, from *t*-1 to *t*.

Call any preference that makes up an agent's genuine evaluative perspective a "welfare-relevant" preference. Given everything that has been said, it appears that the project of solving the status-quo bias is identical to the project of adequately characterizing preferences that are genuinely welfare relevant. There have been several attempts to capture an agent's welfare-relevant preferences. In the next section, I consider two that fail for an illuminating reason.

3. Autonomy and the Ideal Advisor

One important account of welfare-relevant preferences is offered by, among others, Jon Elster and L. W. Sumner. Sumner writes:

> Why are we reluctant to take at face value the life satisfaction reported by 'the hopeless beggar, the precarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie'? Presumably because we suspect that the standards which their self-assessments reflect have been artificially lowered or distorted by processes of indoctrination or exploitation. In that case, the obvious remedy is to correct for the conditions under which their expectations about themselves came to be formed. The problem is not that their values are objectively mistaken but that they have never had the opportunity to form their own values at all. They do not lack enlightenment, or insight into the Platonic form of the good; they lack autonomy.[22]

Importantly, Sumner discusses autonomous "happiness" rather than preferences—I leave this consideration aside given that the proposal, if successful, would succeed for preferences as well as for "happiness". For Sumner, the status-quo bias is a result of non-autonomous mechanisms, both social and psychological, that produce biased happiness or biased desires and preferences.

[22] Sumner, 166.

On this view, preferences that are formed by a process that is not compatible with an agent's autonomy are rejected as welfare-irrelevant, preferences that are—or would be—formed by processes that are compatible with an agent's autonomy are treated as expressing the agent's genuine evaluations, and are hence accepted as welfare-relevant. (Call this the "Sumnerian account".)

Though Sumner doesn't commit to any particular theory of autonomy, he does commit to the non-autonomous status of certain preference (or happiness) formation processes.

> Self-assessments of happiness or life satisfaction are suspect (as measures of well-being) when there is good reason to suspect that they have been influenced by autonomy-subverting mechanisms of social conditioning, such as indoctrination, programming, brainwashing, role scripting, and the like. […] [T]he best strategy here is to treat subjects' reports of their level of life satisfaction as defeasible—that is, as authoritative unless there is evidence that they are non-autonomous.[23]

It is likely that such socialization processes can give rise to a status-quo bias. One might imagine, for instance, that the landless laborer's life of poverty is an importantly non-autonomous preference formation process. But the status-quo bias cannot be fully captured by claiming that desires or reports of satisfaction or happiness are non-autonomous, in the way spelled out by Sumner. Indeed, if anything is an "autonomous" process of preference revision, Erin's surely is.[24] Erin prefers the life of a dancer for its own sake, but finds that she has no talent for it. She then, quite rationally, decides to stop pining away for the life she longed for, and strategically attempts to revise preferences toward the available: the life of a painter. Indeed, far from being an instance of pernicious socialization, embarking on a strategy of preference revision of this kind is straightforwardly rational. Why continue to pine away for something when it is clearly beyond your grasp? Preference switching in this case will lead to far fewer cases of frustration and regret which is surely, all things considered, something to be praised and encouraged.[25]

---

[23] Sumner, 171.
[24] Compare Rickard, 289-90.
[25] Cf. Connie Rosati, "Preference Formation and Personal Good" in Olsaretti, op cit., esp. 60-64.

My argument here relies on a sensible principle of preference revision. Consider

> *Anti-frustration*: when an agent cannot achieve some object *x* that she prefers, it is *prima facie* rational for that agent to try to revise her preferences away from *x*.

At least in most cases, when an agent cannot achieve some object *x* that she prefers, it is *prima facie* rational for her to attempt to revise her preferences away from those things she cannot achieve. In many cases, an agent will have decisive reason to do so.[26] As Nussbaum writes:

> someone as a child may want to be the best opera singer in the world (as I did), or the best basketball player—but most people adjust their aspirations to what they can actually achieve. ... I am not free to be a leading opera singer, nor is a short adult free to be a leading basketball player. We have failed to reach the grapes, and we have shifted our preferences in keeping with that failure, judging that such lives are not for us. But clearly this is often a good thing, and we probably shouldn't encourage people to persist in unrealistic aspirations.[27]

Insofar as frustration and regret are bad and Erin has reason to avoid them, and insofar as continuing to prefer dance to painting will hinder her achievement of a good life, it is irrational to refuse to at least make an attempt at preference revision.[28] In order for the autonomy constraint to avoid the status-quo bias in the face of this seemingly sensible principle, it must be the case that rational preference revision of this kind is non-autonomous.[29] But this seems wrong. Any agent that can't accomplish some goal will have reason to revise her preferences downward. Surely we don't want to say that *omnipotence* is required for autonomous preference formation.

One might reply that if they are the result of a rational process of preference revision, we have no reason to believe that biased preferences are not a guide to an agent's good. This

---

[26] One thing that might come between the *prima facie* and *genuine* rationality of this kind of preference revision might be an instrumental benefit of not revising a preference, in terms of overall preference fulfillment. For example, one might prefer something that cannot be achieved, but so preferring might cause someone to fulfill other preferences that would not have been fulfilled otherwise. In addition, you might think of cases in which the unachievable preference is, say, central to this person's identity (such that giving up the preference would be repugnant in itself, despite its guaranteed frustration). In this case there may not be genuinely rational, despite its *prima facie* rationality.

[27] Nussbaum, *Women and Human Development*, 137-8.

[28] I here remain neutral concerning whether one can ever have reasons *for* preferences themselves. At the very least, Erin certainly has a reason to try, or attempt, or seek to prefer the life of hedonic satisfaction, and hence the process that brings about such preference revision can be perfectly autonomous. See Derek Parfit, "Rationality and Reasons" in *Exploring Pratical Philosophy: From Action to Values*, ed. Egonsson, Josefsson, Petterson, and Rønnow-Rasmussen (Aldershot: Ashgate, 2001).

[29] Elster seems to endorse a claim like this, which Nussbaum rightly criticizes. See Elster, "Sour Grapes: Utilitarianism and the Genesis of Wants," 228-9, and Nussbaum, *Women and Human Development*, 138.

suggestion is mistaken. Merely because it is *rational* to at least attempt to switch her preferences does not mean that the life of a dancer wouldn't be good for Erin. If she *could* achieve it, we are tempted to say, it would certainly be good for her. *The Painter* is a problematic case in precisely the way the status-quo bias is problematic—by violating *Independence*, and hence interrupting the inference from one's preferences to one's good, not because the process by which the preferences were formed is somehow irrational or non-autonomous.

Another account of welfare-relevant preferences is worth comment here. Some hold that an agent's genuine evaluative perspective is not given by her actual preferences, but by her *second-order* preferences, or preferences that are held on her behalf by an "ideal advisor". For instance, according to Peter Railton, an agent's good is determined by what a second-order agent, suitably informed, would want the first-order agent to want: "The proposal I would make, then, is the following: an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error and lapses of instrumental rationality."[30] David Lewis proposes a similar "dispositional" view: "to be a value—to be good, near enough—means to be that which we are disposed, under ideal conditions, to desire to desire."[31] Lewis' ideal conditions will include full imaginative acquaintance with all possibilities. Call these—importantly distinct—views "second-order accounts" of welfare-relevant preferences.

When it comes to the status-quo bias, second-order accounts do no better—and indeed appear to do much worse—than the Sumnerian account. For instance, at *t*-1, Happy does not yet know that he possesses no talent for hockey. Hence he hasn't yet adapted his preferences to that which he could achieve. But—taking Railton's view for the moment—an ideal advisor, knowing that Happy is talentless, will—even at *t*-1—*want Happy to want* hedonic satisfaction rather than

[30] Railton, 54.
[31] Lewis, 71.

hockey achievement for its own sake. More generally, it is unclear that a fully informed and fully rational advisor would want a first-order agent to develop *any* desires that could not be satisfied or have a low chance of satisfaction. Recall *Anti-frustration*. When someone prefers that which cannot be achieved to that which can be achieved, it is rational—at least *prima facie*—to revise that preference in light of the inability to achieve the preference's object. And if so, because success as a hockey player is unavailable to Happy, the ideal advisor will not want him to prefer hockey playing to the life of hedonic satisfaction. But if this is the case, the ideal advisor view appears to imply that the life of hedonic satisfaction is better for Happy even prior to his preference shift.

The problem is similar for Lewis' dispositional view. With full imaginative acquaintance of a life of great hockey achievement, Happy might come to desire it more than hedonic satisfaction. But would he *desire to desire* it (or desire to prefer it over hedonic satisfaction), knowing that it cannot be achieved? Plausibly, no.[32] But then, on the dispositional view, hockey achievement appears to be worse for Happy than hedonic satisfaction (given that the latter is likely to be the object of a second-order desire). Hence, the status-quo bias.

For ideal advisor views, the status-quo bias crops up even in cases in which *Anti-frustration* is not in play. Consider, for instance, the following case. At *t*, Charley is in prison and prefers, for its own sake, to be out of prison. Unfortunately, as a criminal, Charley is an incompetent stooge. Even more unfortunately, Charley always attempts to achieve those things he desires. Given Charley's incompetence, his prison break will fail and he will be sent to solitary confinement, which he hates more than mere prison. However, Charley's cellmate Clint

[32] Compare the following from Parfit, "Rationality and Reasons", 27: "If we believe that our having some desire would have good effects, what that belief makes rational is not the desire itself, but our wanting and trying to have it." As Parfit notes, an ideal advisor can *want* a first-order agent to want to desire some particular object because not desiring it, or failing to desire it, will have no good effects when it comes to the rest of the agent's desire set. Indeed, if the ideal advisor is rational, he will want a first-order agent to want those desires that will have the best overall effects. But these desires, as shown here, will result in preferences that are biased toward the status-quo.

is a first-class criminal, and has been working in secret on his own prison break. Were Charley to hang back, Clint would spring Charley at *t*+1 without any effort on Charley's part, despite Charley's incompetence. Surely, we wish to say, getting out of prison is better for Charley than staying *in* prison. But given that Charley attempts everything he desires, a second-order agent, fully informed, would *want Charley to want* to stay in prison, at least until *t*+1. Were Charley under ideal conditions, he would not desire to desire to escape. But then are we to say that it is *better* for Charley to be in prison than out of prison at *t*? Surely not!

The problem with the second-order views is shared by the Sumnerian account. We appear to have reason not to shoot for the stars. Hence what is in our best interests is one thing, but what we desire to desire, or what we might come to autonomously prefer, is another thing altogether. Any view that links an agent's good to such second-order or autonomous processes will—at least on occasion—violate *Independence*.

I think the intuition I noted in the previous section is correct. The status-quo bias appears to interrupt an inference from an agent's preferences to the agent's genuine evaluative perspective. In the next section, I sketch an account of welfare-relevant preferences that solves the status-quo bias and hence issues preferences that plausibly form an agent's genuine evaluative perspective.

4. Preference Coherentism

Call my account of welfare-relevant preferences "preference coherentism" (PC). I won't be able to defend all the nooks and crannies of PC here; I set out only to show how it might solve the status-quo bias. There may be other good reasons to reject it; I leave these aside. (For instance, I will not defend my precise conception of coherence here.[33] Furthermore, some have argued that my account is ultimately circular, I dispute this elsewhere; here I confine this

[33] I try to specify a working notion of coherence good enough for our purposes here in "A Coherence Theory of Truth in Ethics" in *Philosophical Studies* 117 (2006).

discussion to note 34.[34]) PC declares that an agent's welfare-relevant preferences, and hence her genuine values, are constituted by her *conception of the good, rendered coherent and complete*. I render this account in three stages. First, I argue for my understanding of a conception of the good. Second, I more fully describe the "coherence and completeness" constraints. Third, I argue that PC solves the status-quo bias.

4.1. Beliefs not Desires

The first noteworthy feature of preference coherentism is that it understands preferences in terms of evaluative beliefs rather than desires. *A*'s welfare relevant preference for *x* over *y* will, for PC, take the form of a belief that *x* is better for *a* than *y*. This requires defense. Most importantly, the appeal to beliefs rather than desires in defining welfare-relevant preferences solves at least one way in which status-quo biased preferences might arise. This proposal is intuitive, as noted by Sidgwick: "a prudent man is accustomed to suppress, with more or less success, desires for what he regards as out of his power to attain by voluntary action—as fine weather, perfect health, great wealth or fame, etc.; but any success he may have in diminishing the actual intensity of such desires has no effect in leading him to judge the objects desired less

[34] See, for instance, Lewis, 70; Brink, 20-21. Why might this view be circular? If—at least occasionally—well-being is explained in terms of beliefs about well-being, and the *content* of beliefs about welfare appears to involve well-being, then the *semantics* of judgments about well-being, or what would be good for me, are circular. But this is not so. The solution here lies at the level of semantics: though I shall remain substantially uncommitted on various possibilities, one can give an account of the content of beliefs about well-being that does not involve the agent's own judgments. In fact, one can give any such semantics. For instance, one might claim that one's judgments about well-being are judgments about some non-natural property, say, *property x*. Whether this property actually exists is neither here nor there. On my view, judgments about well-being permit of a coherence, rather than correspondence, theory of truth. Whether *property x* exists, or is properly predicated in such-and-such a way, is thus irrelevant in determining the truth of judgments about well-being. This does not require the rejection of objective prudential values: one might insist that the value of coherence as a truth-maker is confined in certain ways by the presence or absence of external facts about value, depending on one's first-order theory of well-being. Though the eventual account might result in a complex theory of truth as applies to judgments of well-being (sometimes correspondence, sometimes coherence), it is only as complex as is required given the most plausible theory of the good. This is as it should be. See Dale Dorsey, "Subjectivism without Desire", MS, and "Truth and Error in Morality", forthcoming in *New Waves in Truth*, ed. Wright and Pedersen. I thank Doug Portmore for forcing me to address this important objection.

'good.'"[35]

This intuition can be made a bit more concrete by considering an example of Jennifer Hawkins'. Hawkins, in discussing two characters who develop low self-esteem or self-worth and come to prefer their intuitively terrible conditions in which they live, writes as follows:

> One particularly striking feature of both Celie and Quoyle is their low sense of self-worth. As I suggested before…Celie *feels* worthless. Without a proper sense that she, Celie, matters, her idealized self will be unlikely to give her non-idealized self good advice. Nor will she be able to assess her life fairly in comparison with other lives. Although she may recognize that other lives are happier or more successful, she will not necessarily see them as better for her if she herself feels unworthy of the lives in question.[36]

I think Hawkins' challenge is important. Some people, because of their circumstances, will develop a lack of self-worth and will, as a result, develop preferences that are biased toward the status-quo. This might occur not only for people whose lives are objectively bad (see the landless laborer), but also in others; one could imagine, for instance, that Erin's preference for painting might arise given a conviction that she is "unworthy" of the life of a dancer. But it is important to note that there are (at least) two ways in which low self-worth could bias someone's preferences. First, a person with low self-worth might *believe* that other lives would be substantially better than her current life, but because she has a low sense of self-worth and "feels unworthy of the lives in question", she fails to desire them, or fails even to want to want them. Second, a person with low self-worth might not believe that she is unworthy of that which she believes is good, but might simply come to judge that the life in which she is, say, dominated, landless, overexhausted is actually good for her, perhaps because it somehow "fits" her low conception of her own worth.

I will leave discussion of the second possibility until section 4.4. But PC's emphasis on beliefs rather than desires helps to address the first permutation of Hawkins' analysis: even though Celie might believe herself unworthy of that which she believes is good, she still has a

[35] Sidgwick, 110. Sidgwick's own account, however, fails to properly address this problem: Sidgwick himself insists that a person's desires are authoritative over his good only among those choices that are "open to him". See Sidgwick, 111-12.
[36] Hawkins, 160-1.

conception of what is good for her—she *believes* that lives other than her own would be better than the one of which she feels herself unworthy. Her lack of self-worth causes a status-quo bias only at the level of desire: because she lacks self-esteem, she fails to want that which she judges good for her. Though this might not be the only way in which low self-worth can contribute to the status-quo bias, it is certainly *one* way it might do so. Hence understanding welfare-relevant preferences in terms of evaluative beliefs rather than desires helps to correct this status-quo bias, and is one reason to treat a person's evaluative beliefs as her genuine evaluative perspective.

4.2. Coherence and Completeness

Understanding an agent's preferences in terms of evaluative belief rather than desire goes some distance toward solving the status-quo bias. But it does not go far enough. An account of welfare-relevant preferences must be abstracted from the agent's actual conception of the good, and must be identified with what a person *would* prefer under certain counterfactual conditions. There are two reasons for this. First, a given agent's conception of the good might be incoherent, inconsistent, or self-refuting. Second, without such abstraction we cannot solve the status-quo bias (even if we reject a desiderative interpretation of preferences). Beliefs, no less than desires, can be subject to such biases, through any number of psychological mechanisms (including rational revision *a la Anti-frustration*). As we have seen, however, the form of abstraction one selects is of the first importance: the second-order view defines one's well-being in terms of that which would be preferred in certain counterfactual conditions, but in a way that only worsens the status-quo bias. I select two forms of abstraction here.

First, conceptions of the good must be understood not as that which any agent happens to believe as good, but as that which an agent *would* believe good *were* the agent's actual conception of the good rendered *coherent*. Incoherence involves not only contradictory beliefs, but beliefs that are ill-behaved in various ways, including those that display intransitivity.[37]

[37] Though transitivity has been denied as a feature of the good, I accept it as bedrock. For conflicting

Furthermore, coherent beliefs support each other and provide explanatory and justificatory connections. Though I take this to be a rather weak requirement, we should at least insist that welfare-relevant preferences that purport to reflect an agent's genuine evaluative perspective be supported and warranted by other beliefs that are members of the agent's conception of the good.[38] Hence PC holds that *a*'s preference for *x* over *y* is welfare-relevant only if this preference is part of *a*'s conception of the good were *a*'s conception of the good coherent or, if *a*'s conception of the good is not coherent, *a*'s conception of the good *rendered* coherent. (More on this below.)

Second, welfare-relevant preferences should be construed as arising from a *complete* conception of the good. The completeness requirement is really two requirements in one. The first is that a welfare-relevant conception of the good must yield a *complete ordering* of all possible welfare goods. (If commensurability is limited, an agent's coherent and complete conception of the good should yield *as complete an ordering as possible*.) By "possible", I mean *metaphysically possible*. Insofar as we believe that a change in an agent's capacities might be good or bad for *a*, we shouldn't hold these fixed when issuing a complete account of that which would be better or worse for *a*. This is a crucial step in avoiding the status-quo bias: if one's conception of the good can rank-order activities or other welfare goods that are possible only in some more restricted sense of "possible", this will violate *Independence*.

However, in order to guarantee such an ordering, a conception of the good must have a sufficient basis to *determine* a complete ordering without gaps. This consideration naturally leads into the second half of the "completeness" requirement. The agent's conception of the good must

---

views, see Stuart Rachels, "Counterexamples to the Transitivity of 'Better Than'" in *Australasian Journal of Philosophy* 76 (1998); Larry Temkin, "A Continuum Argument for Intransitivity" in *Philosophy and Public Affairs* 25 (1996). Arguments, in my view successful, for transitivity are to be found in John Broome, *Weighing Lives* (Oxford: Oxford University Press, 2004), ch. 4; Alistair Norcross, "Contractualism and Aggregation" in *Social Theory and Practice* 28 (2002).

[38] I argue for this in more detail in "A Coherence Theory of Truth in Ethics", op. cit; and "Subjectivism without Desire", MS.

be tested against a complete set of *value data*. Complete testing closes potential gaps. Though I now have no beliefs that will yield a proper ordering between the life of an Aztec chieftain or a Mayan chieftain, testing my conception of the good against the relevant value data will close this gap (or, if they are genuinely incommensurable, declare them so). Furthermore, the completeness constraint mirrors a general virtue of webs of belief. A web of belief is more trustworthy when it has been tested against more data. A web of belief is maximally trustworthy when it has been tested against all possible data. Unless we are influenced by a pragmatist account of truth, we wouldn't say that this entails that the web of belief is true in the *scientific* case. However, we would (or should) say that when a conception of the good has been tested against all possible value data, it *determines* an agent's welfare-relevant preferences. Thus the completeness requirement refers both to the mandated complete ordering, and also to the requirement that any welfare-relevant conception of the good survive a counterfactual process of complete testing.

What are value data? On my understanding, a value datum consists of two crucial elements. First, the *information* about *what living a given life would be like*. A value datum will require the full confrontation with the consequences, experiences, achievements, etc., of living some life. There are different ways one might construe this element of value data. One might conceive of it as a "report", i.e., some list of facts about the content of a life, or as an "experience", i.e., the actual *experience* of living a life reported on in the report model.[39] According to David Sobel, the choice is not inconsequential: a mere report, in comparison to the full experience, will fail to accurately convey the bases for a complete evaluation.[40] Sobel's critique seems correct, and hence I will accept it for my purposes here. Thus value data will require full information about what a given life would be like for the person who lives it, conceived of as the experience of actually living it. The second element of value data is the *judgment* about the quality of that life *given* the experience of it. Thus value data is properly

[39] David Sobel, "Full Information Theories of Well-Being" in *Ethics* 104 (1994).
[40] Sobel, 798.

conceived of as a belief in the quality of some particular life on the basis of actually experiencing that life, *and on that basis only*.

One point remains. The coherence and completeness requirements introduce the possibility of recalcitrant data: a belief about the good that conflicts, or renders incoherent, the conception of the good it is used to test. If so, revisions to the conception of the good with an eye toward renewed coherence will be required. How are we to go about rendering coherent an incoherent set? Here PC remains conservative. When revising a conception of the good in light of recalcitrant value data or as a result of incoherence, revisions are made at the "periphery", revising an agent's most strongly held preferences only as a last resort. Thus, putting all this together, PC holds that welfare-relevant preferences are preferences that make up the agent's conception of the good, after having been tested against all possible value data, and revised in light of recalcitrant data and other forms of incoherence by a process of "minimal mutilation".[41]

4.3. Coherentism and the Status-Quo Bias

PC solves the status-quo bias. Because PC holds that welfare-relevant preferences are drawn from a complete ordering of all possible lives, and because that ordering is generated by a conception of the good that is completely tested against judgments about the value of lives made on the basis of experiencing those lives, there is no room for an agent's good to be more heavily weighted toward the one she actually lives. Her actual life will of course be evaluated on the basis of the experience of actually living it. But this value datum will be only one among many value data that will be used to construct and test a coherent conception of the good.

The key here is not the fact that preference coherentism corrects false beliefs about alternative lives, though it does of course do that. Rather, PC blocks the cognitive conditions that yield a status-quo bias. The difference between biased preferences and unbiased preferences is the extent to which facts of *availability or unavailability* influence the extent to which the agent

[41] W. V. Quine, "Two Dogmas of Empiricism" in *From a Logical Point of View* (Cambridge, MA: Harvard University Press, 2nd ed., 1981), 42-6.

in question maintains the preference. Status-quo biased preferences arise on the basis of these facts. Were hockey achievement not unavailable, Happy would still prefer it, hence the preference for hedonic satisfaction rather than hockey achievement displays a status-quo bias. But when revising an agent's conception of the good according to preference coherentism, the facts of availability are rendered moot. Though it might be the case that Erin cannot live the life of a dancer, and hence develops a biased aversion to such a life, the *value data* that are used to revise her conception of the good are not influenced by Erin's actual inability. Any relevant value datum used to revise her conception of the good includes a judgment of the comparative value of a particular life and its various goods given on the basis of a full experience of that life, not whether that life is available or is likely to occur, given the facts about the world. Facts of actual availability will not influence the extent to which recalcitrant value data will cause revisions in her conception of the good. Because all lives are experienced, preferences that depend on facts of availability will be—other things equal—revised. On PC, the life that an agent actually leads is simply one among many, and has no special status.

Of course, the status-quo bias is an ineradicable feature of value data: the *experience* of living a life will surely present many *examples* of biased preferences. But in coming up with a coherent and complete conception of the good, the preferences one has during any particular life are not necessarily welfare-relevant. Though, in life *x*, Happy might strongly prefer the hedonic life to hockey, Happy's coherent and complete conception of the good can declare this preference welfare-irrelevant. And it might do so by comparing life *x*'s value datum to life *y*'s value datum: the life in which hockey achievement is available and experienced is likely to be strongly valued, sufficient to override life *x*'s value datum.[42]

It might be objected that preference coherentism is still beholden to the (perhaps biased)

[42] However, were Happy to judge his life as a hockey player worse than his life as a hedonist from the perspective of his coherent and complete conception of the good, being a hedonist *is* a better life for Happy. Furthermore, if the respective value data judge these lives to be equivalent in value, Happy's conception of the good will also reflect this fact. This is the correct answer.

preferences of the original agent in one sense: the agent's conception of the good is revised conservatively, i.e., by minimal mutilation, which requires that we overrule an agent's most strongly held judgments only as a last resort. It is not at all guaranteed that preferences that are originally a result of facts about availability will not be among the agent's strongest held beliefs from the perspective of her conception of the good. After all, over the course of time and psychological change an agent might come to strongly identify with preferences that were originally cases of a status-quo bias, or other forms of psychological distortion. If so, confrontation with value data is insufficient to remove biased preferences because this reflection will surely be done in light of these very preferences. However, PC has resources to alleviate this worry. After all, *in order for the preference to be an example of the status-quo bias*, there must be some reason for thinking that the agent's preferences arise merely given the facts about what is or is not available. Because Erin's biased preferences arise and are supported by facts of unavailability, Erin will prefer achievement in dance when these facts are rendered moot. Similarly for Happy. In experiencing value data, there is no longer any feature of the world—no fact of availability—that would allow biased preferences to develop, or recommend preference revision *a la Anti-frustration*. All beliefs are subject to revision, and given the nature of biased preferences, they are likely to be revised when facts of availability are no longer relevant.

But what about an agent whose psychology was altered such that his once-biased preferences are now—at *t*—*entrenched* parts of his coherent conception of the good (such that they would not be revised by the counterfactual process outlined here)? It seems to me that these cases do not display the status-quo bias. We should not insist that all preference revision, indeed, all revision of a preference toward that which one can obtain, is an instance of bias. Any preference surviving the process endorsed by PC displays precisely the characteristics of unbiased preferences: it is not dependent for its existence upon facts of the availability of one life over another, or on facts about what the status-quo actually is. Given that this is the case, we should

not describe this as a problematic form of status-quo bias, but rather as an important and genuine form of status-quo *value*. After many years as a hedonist (say, at t+1), Happy's conception of the good might be such that, made coherent and complete, hockey is no longer regarded as a valuable pursuit. If so, PC will reflect this change in Happy's conception of the good, and will declare his preference at t+1 for hedonic satisfaction welfare-relevant.

This is the correct answer, and illustrates an important way in which PC can accommodate the original motivation for preferential authority over well-being. In solving the status-quo bias, PC appears to be a plausible theory of an agent's true values. In addition, despite its abstraction, PC does not issue verdicts about well-being that are alien to the agent herself. PC follows conservative standards of revision, and hence will revise values that Happy holds most dear only on the strongest grounds of bias, distortion, or mistake. Thus, though Happy's welfare-relevant preferences will rarely be identical to the ones he now holds, insofar as they are firmly grounded in the values he holds most dear, they should be regarded in no uncertain terms as *his*. Happy's coherent and complete conception of the good, plausibly, *is* his genuine evaluative perspective.

PC also illuminates my response to an objection stated in §2. There it was argued that Erin's t preferences should be treated as authoritative, given her goals shift from dancing to painting at t. But because biased preferences will not form part of a coherent and complete conception of the good, it is false that Erin's *welfare-relevant* preferences shift from t to t-1. Hence neither Happy nor Erin's case is a case of welfare-relevant preference change at all. Given the case as described, Erin's welfare-relevant preference for dancing at t-1 (presuming, of course, that this preference will appear in her conception of the good at t-1) will persist at t.

4.4. Self-Worth Reconsidered

One problem remains. Take again Hawkins' discussion of self-worth. It might be that Celie fails to believe that anything but the horrible conditions in which she lives is *good for her*.

Her lack of self-worth might *result* in a certain belief in the value of the life one currently lives being strongly held, resistant to recalcitrant value data. Perhaps, for instance, the landless laborer has such a low self-worth that his belief that landless labor is better than alternatives would survive the process of complete testing. If so, one might object, PC cannot adequately capture our intuition that no matter what Celie prefers, her self-worth-affected preferences do not determine that which is good for her.

We should distinguish two questions here. First, is Celie's preference welfare-relevant? Second, is Celie's preference authoritative? I think the answer to the first question must be "yes". Celie might, after years of low self-esteem, come to believe that the life she lives, which contains very little happiness, achievement, genuine friendship, or hope, is actually the *best life for her*, in a way that would not be changed by honest confrontation with the *experience of living all possible alternative lives*. If so, it is very difficult to see Celie's preferences as an example of the status-quo bias, or as expressing anything but Celie's genuine values. She values the status-quo, but not in a biased way; not *because* it is the status-quo. Celie's preference is, for this reason, welfare-relevant.

Take now the second question. Intuitively, it seems quite right to say that Celie's current existence is worse for her than almost any alternative. This might be for two reasons. First, Celie's life, like the life of the landless laborer, is simply bad in a way that is independent of preference. Celie's genuine evaluative perspective, in a very real sense, gets it wrong. Second, we might also hold that low self-worth is an objective bad-making feature of lives. Hence one might say that even someone who does not prefer the worse but who prefers his current life only as a result of poor self-worth nevertheless lives a worse life than he otherwise would, *given* his lack of self-worth.

But preference coherentism is compatible with these sensible suggestions. Merely because Celie's preference is welfare-relevant does not entail that Celie's preference is

authoritative. Though Celie's preference might reflect her genuine evaluative perspective, she clearly *prefers the worse*. Furthermore, it seems plausible to say that her life is made worse by her overall lack of self-worth. As I argue above, welfare-relevant preferences ought to have at least some authority when it comes to well-being. Just how much authority they have depends on a range of factors including the extent to which these preferences conflict with or do not take account of weighty preference-independent facts about well-being. PC is intended not to be a complete account of welfare, but rather a theory of welfare-relevant preferences. Insofar as PC has a plausible claim to represent Celie's genuine, unbiased evaluative perspective, it succeeds—despite a conviction that preferences do not, by themselves, settle all facts about welfare.

5. Conclusion

In this paper, I have argued that the status-quo bias is a substantial problem for any plausible theory of well-being. An agent's values seem indispensible in determining that which is good for her. Previous attempts to avoid the status-quo bias fail or, in some cases, make the problem substantially worse. I have also argued that where previous attempts to contain the status-quo bias have failed, preference coherentism succeeds. PC is not a complete theory of welfare. The fulfillment of *a*'s welfare-relevant preferences is not sufficient to determine that which is good for *a* at *t*. However, given the plausible rationale for consulting preferences in determining that which would be good for *a* at *t*, an examination of *a*'s complete and coherent conception of the good appears to be a necessary feature of such a determination.