Adaptation, Autonomy, and Authority

Dale Dorsey

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

Adaptation is a fact of life. If I walk down the street and fall down a manhole, I'm likely to keep a lookout for open manholes in the future. In this case, I *adapt* to my circumstances; I take a more cautious approach than I might otherwise have. This is true of all forms of life, and plausibly helps to explain the fact that there is, in fact, life at all.

Adaptation of *preferences* is not much different. If I have a preference for something that I can't get, I'm likely to adapt: I will either stop preferring that thing, or perhaps (depending on the form my adaptation takes) come to reverse my preferences. If I desire to go to Harvard and not Columbia, but I get into Columbia and not Harvard, I may come to revise my preferences toward Columbia, and away from Harvard. There could be many reasons for this adaptation, but one chief reason is that to maintain the former preference is painful: it is the frustration of something I want. Perhaps not as painful as falling down a manhole, but a pain no less worthy of psychological adaptation to avoid.

This is a simple fact of human life, unremarkable except for the fact that it causes widespread—and widely noted—problems in moral theory. Preferences are supposed to represent, broadly speaking, a person's *good*; they are supposed to represent the locus of a person's *autonomous decisionmaking*; they are the object of our beneficent concern; and, perhaps somewhat more controversially, the proper index by which to measure social choice. But if preferences can be adaptive in this way, each of these roles appears to disintegrate. It would be wrong, for instance, to say that I'm doing better to the extent that I'm going to Columbia rather than Harvard, that my choice of Columbia rather than Harvard was somehow autonomous, that the person with a beneficent interest in my welfare would promote a state of affairs in which I attend Columbia *rather than* Harvard, or that a proper social theory should design institutions or policies with an eye toward this preference. The problem gets worse, of course, when individuals adapt to

conditions that social policy should generally try to avoid, such as poverty or oppression.

The existence of adaptive preferences might cause one to believe that preferences should play no role in normative domains to which they may be applied.[1] But this inference is too quick. After all, in cases in which a person's preferences are *not* adaptive, it would seem implausible to hold that proper accounts of autonomy, welfare, beneficence, etc., should make no reference to an individual's preferences. But if this is right, a crucial question for ethics and social theory is: is there a principled method by which to distinguish normatively authoritative preferences, and to expunge those that are problematically adaptive?

In this essay, I investigate the connection between adaptation and normative authority. However, my investigation into the concept of normative authority reveals something of a puzzle. Though there is good reason to believe that adaptive preferences are, broadly speaking, failures of *autonomy*, I conclude that there is no account of preferential autonomy that can plausibly eliminate all forms of preference adaptation. Hence, I offer two potential solutions to this puzzle: first, that we should simply reject the claim that preferences should play a normative role (insofar as there is no acceptable method by which to distinguish the normatively authoritative ones from those that lack normative authority), or, second, that we should reconsider whether all forms of preferential adaptation signal a lack of normative authority. Briefly, I argue in favor of the latter, and hopefully more optimistic, conclusion.

1. Varieties of (the Lack of) Normative Authority

It seems right to say that the fact that a particular preference for ϕ rather than ψ is adaptive is a reason to look askance at its normative authority. We generally refuse to treat it as a guide to a person's welfare, refuse to take it seriously in planning our beneficent actions, and hold that social choices should not be indexed to the fulfillment of such a preference, etc. (Whether this is actually true will turn out to be a relatively complicated matter to which I will return at the end of the paper.)

But I should distinguish the problem of normative authority that results from problems of adaptation from the problem of normative authority that

[1] See, for instance, Martha Nussbaum, *Women and Human Development* (Cambridge: Cambridge University Press, 2000), ch. 2 for a critique of preferences rather than *capabilities* as the proper index of social choice. I criticize Nussbaum's position in *The Basic Minimum: A Welfarist Approach* (Cambridge: Cambridge University Press, 2012), 19-32.

results from other sorts of preference failures. Adaptation is not the only feature of a preference that generates a lack of normative authority. Preferences can lack normative authority because they are *sadistic*.[2] If I prefer to harm you, for instance, it might be that this preference displays a normative failure. In addition, preferences can lack normative authority because they are *trivial* or, indeed, directed toward *the objectively worse*.[3] Not all agree that such preferences lack authority. But we should admit that adaptation *per se* should be distinguished from failures of authority for such reasons. I can maintain a sadistic preference non-adaptively. I can adaptively prefer something that is sufficiently objectively good, etc.

For the purposes of this paper, I want to focus on problems of normative authority that stem specifically from adaptation. This is not to say the remaining problems are not particularly serious. But they are, I think, important to keep distinct.

2. Adaptation and Autonomy

Then what distinguishes the *per se* failure of normative authority associated with adaptation? What, more to the point, *explains* why adaptive preferences are not authoritative (in the way that, say, a bad will explains why sadistic preferences are not authoritative, or the objective good explains why trivial preferences lack authority)? I think the right answer,[4] appeals to the inapplicability of a central rationale for taking preferences seriously at all. Insofar as we care about preferences in moral and political theory, we care about them because they seem to capture what people *value*, or what expresses their own evaluative point of view. We believe that the fact that someone values something has normative or evaluative consequences: it *makes* the thing valued valuable, or *pro tanto* worth pursuing. But sometimes, as in cases of adaptation, preferences do not genuinely express what someone values. They do not, in Sumner's terms, "reflect the subject's own

[2]See, for instance, John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), 30-1.

[3]See, for instance, David Brink, "The Significance of Desire" in *Oxford Studies in Metaethics* v. 3, ed. Shafer-Landau (Oxford: Oxford University Press, 2008), 24-25; Richard Kraut, "Desire and the Human Good" in *Proceedings and Addresses of the American Philosophical Association* 68 (1994), 41-42.

[4]And, indeed, the most common answer. See, for instance, L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Oxford University Press, 1996), ch. 6; Jon Elster, "Sour Grapes: Utilitarianism and the Genesis of Wants" in *Utilitarianism and Beyond*, ed. Sen and Williams (Cambridge: Cambridge University Press, 1981), 226-230.

point of view".[5] In other words, adaptive preferences display a failure of *autonomy*—a failure to express what *they* really value.[6] Though this idea is not precisely formed (an investigation into which, I should note, forms much of the remaining content of this essay), it seems right to say that when I adapt my preference to Columbia and away from Harvard, this is an adaptation *away* from my genuine values, or an attitude that accurately expresses my autonomous point of view. It is, as it were, putting on a kind of "mask": adopting a preference or evaluative attitude that does not reflect *me*. That adaptive preferences are non-autonomous seems essential to the concept and function of an adaptive preference. Adaptive preferences are *adaptations*: they alter the preferences we have in light of external circumstances, in particular, facts about the way the world is, or facts about what is available or unavailable to us.[7] Adaptive preferences block one's own genuine evaluative attitudes insofar as maintaining our genuine attitudes, *given* facts about the way the world is, is worse; it is painful, frustrating, or otherwise disadvantageous. And hence it would appear that the essential fact of adaptive preferences, the fact that renders them a phenomenon *at all*, is that these preferences interrupt or mask our own genuine point of view. They are, for this reason, failures of autonomy.

There's another way to see this point. The normative failure of sadistic or shallow preferences is a matter of correspondence of a preference with some external measure, viz., moral demands, the requirements of respect for others, or some (independent of preference) measure of the objective good. But *adaptation* seems different: adaptive preferences do not (or do not necessarily) fail to conform to some external measure, but rather with an *internal* measure. But what is this internal measure? Surely adaptive preferences do not lack normative authority because they do not measure up to what I *previously* valued; this would render virtually all instances of changed preferences normative failures. Rather, adaptive preferences fail to measure up to the index of my genuine—autonomous—preferences or states of valuing. This seems to match up with, as it might be called, the phenomenology of such preferences. When we confront those whose preferences are adaptive, we have a tendency to think that their preferences do not really express what they genuinely value, or would value under conditions

[5] Sumner, 172.

[6] "Why are we reluctant to take at face value the life satisfaction reported by 'the hopeless beggar, the prevarious landless labourer, the dominated housewife, the hardened unemployed or the over-exhausted coolie'?... They do not lack enlightenment, or insight into the Platonic form of the good; they lack autonomy," (Sumner, 166).

[7] Cf. Elster, op. cit.

appropriate to developing autonomous preferences.

The connection between adaptation, autonomy, and normative authority forms the central question of this paper. In essence, it is this: though it is plausible to explain the normative failure of adaptive preferences *via* their lack of autonomy, is there any acceptable account of an autonomous preference that could form a principled method by which to distinguish preferences that lack normative authority as a result of adaptation? One way to investigate this question would be to argue for a particular account of autonomy, independently of that account's ability to offer a plausible explanation of the normative authority of preferences, and only then investigate whether autonomous preferences are normatively authoritative, or fail to be adaptive in a way that causes problems for ethics and social choice. Though this is perfectly open, this is not the style of argument I adopt here. Rather, I want to canvass a variety of potential accounts of the nature of autonomous preferences. I conclude that no such account is acceptable. By way of a conclusion, I assess what this verdict might mean for any plan to take preferences seriously in any normative domain.

3. Autonomy and Autonomous Preferences

As is noted by many theorists of autonomy, this concept is stretched thin. Nomy Arpaly identifies no less than *eight* concepts to which the term "autonomy" can and has been used to refer.[8] Adding to the difficulty here is that autonomy is generally a predicate applied not to preferences *per se*, but rather choices, decisions, or actions. For instance, we may say that an autonomous choice was a choice made without external interference, or a choice made on the basis of reasons.[9] Thus the point at issue is what it might mean to ascribe autonomy to a *preference*. What might it mean to say, for instance, that my preference for Diet Coke over Diet Pepsi is autonomous? Or not autonomous? What *property* are we identifying?

I think accounts of autonomous preferences can be classed into roughly two categories. One might claim that the autonomy of a preference is an *historical* property: a property possessed by a particular preference in virtue of that preference's history or provenance; in particular, the way it was developed or instilled. Or it could be a *time-slice* property, a property that holds, or doesn't hold, of a particular preference at a particular time,

[8] Nomy Arpaly, *Unprincipled Virtue* (Oxford: Oxford University Press, 2003), 118, ch. 4 *passim*.

[9] George Sher, *Beyond Neutrality: Perfectionism and Politics* (Cambridge: Cambridge University Press, 1997), 48-51.

regardless of that preference's history.

On the *historical* side, one might construe a preference as autonomous to the extent that it developed *via* the right sort of process. The question, then, is to identify the right from the wrong processes. I can think of a number of potential accounts. First:

> *Historical Account One* (HA1): A preference is autonomous if and only if the agent in question engages in the process by which it is instilled or developed on the basis of reasons.

Second:

> *Historical Account Two* (HA2): A preference is autonomous if and only if the agent in question would endorse (perhaps under conditions of idealized reflection) the process by which it is instilled or developed.

Third:

> *Historical Account Three* (HA3): A preference is autonomous if and only if the process by which it is instilled or developed gives rise to preferences on the basis of reasons.

Fourth:

> *Historical Account Four* (HA4): A preference is autonomous if and only if the process by which it is instilled or developed is found on an "objective list" of right processes.

In addition to the historical accounts, there are a number of potential time-slice properties that could, in principle constitute the nature of of autonomous preferences. First:

> *Time-Slice Account One* (TA1): A preference is autonomous to the extent that this preference is endorsed by the agent in question (perhaps under conditions of idealized reflection).

Second:

> *Time-Slice Account Two* (TA2): A preference is autonomous to the extent that the agent in question maintains or possesses this preference on the basis of reasons.

Third:

> *Time-Slice Account Three* (TA3): A preference is autonomous to the extent a person has control over whether or not he or she maintains it.

Finally:

> *Time-Slice Account Four* (TA4): A preference is autonomous if and only if its object is endorsed under conditions of idealized reflection.

Just like the historical accounts, there could be additional time-slice accounts, but for the purposes of this paper, I'll limit my investigation to these.

I once again stress that I do not wish to get into a discussion of the exact nature of autonomy. Suffice it to say that each of these proposals seems to capture, at least in some way, the extent to which a particular preference expresses genuine states of valuing (away from which adaptive preferences are adaptations). To see this in more detail, take the historical accounts. Each historical account is tied together by a general thought that adaptation is, after all, a result of problematic *processes*: processes that "mask" an individual's genuine values.[10] Of course, what this amounts to is controversial, and is answered in different ways by each historical account. HA1-HA3 reflect the fact that many people think it plausible to believe that the concept of autonomy is tied very closely to the concept of acting for reasons, or at least acting for what one believes to be reasons. And so it may be plausible to say—as is reflected by HA1—that if a particular preference is the result of a process I engage as a result of reasons that I myself recognize (whether or not these are genuine reasons for action), this process generates autonomous preferences. To describe the right processes in this way is similar to some accounts of the nature of autonomous choice.[11]

Similar thoughts motivate HA2. Let's say that I develop some preference for a particular object ϕ over some other object ψ, and that this preference was the result of some process P. But if we imagine, for instance, that I *endorse* P, perhaps for reasons, or perhaps under conditions of idealized reflection, then we might regard the preference in question as autonomous. (Notice that I construe the "endorsement" suggestion as backward-looking. That is, a preference is autonomous to the extent that the process by which it was instilled is endorsed *at the time at which the preference is maintained*, rather than the time during which the relevant process was ongoing.[12]) It

[10] This thought is well captured by Elster, 226-8.

[11] Cf. Sher, op. cit.

[12] Thanks to the editors of this volume for helpfully articulating this ambiguity.

developed in a way that I regard as reflecting, say, *my own values*. HA2 differs from HA1, insofar as there is no constraint on HA2 that suggests the process must *actually* be engaged on the basis of reasons. It is enough to say, perhaps after the fact, that the process itself was valid, valuable, or otherwise endorsed by the person in question.

HA3 retains the broad connection between autonomous preferences and reasons, but draws this connection in a slightly different way. HA3 holds that a process is autonomous if that process itself includes coming to preferences *on the basis of reasons*. For instance, imagine that I prefer to refrain from smoking rather than to smoke. If I came to that preference on the basis of becoming educated about the health effects of smoking, and thus saw reason to avoid smoking and developed preferences on this basis, this process is of the right sort and hence my preference is autonomous. But if my preference not to smoke was simply the product of disgusting images on the front of cigarette packages, which did not engage my capacity to recognize reasons, this process is not autonomous. HA3 guarantees that autonomous preferences are those that I develop on the basis of reasons I recognize. And if so, it seems plausible to say that such preferences will, broadly speaking, reflect my genuine conception of the good.

Finally, HA4 rejects the possibility of coming up with a single, unitary conception of the nature of processes of the "right sort". Akin to the "objective list" theory of well-being, which holds that a person's life goes well to the extent that it manifests particular items on a pre-determined list, this proposal says that processes generate autonomous preferences just in case such processes are identified on an objective list of autonomy-generating processes. Or, perhaps, that are *not* on an objective list of *wrong* processes; this list might include processes such as oppression, lack of opportunity, poverty, brainwashing, etc. Indeed, just this sort of view is floated by Sumner (in discussing the nature of autonomous life satisfaction). In despairing of the possibility of coming up with an adequate unifying account of the nature of autonomous processes, he writes instead that:

> It appears, therefore, that neither of the currently dominant theories about the nature of autonomy is self-sufficient... However the details of a fully adequate view are worked out in the end, the implications for our theory of welfare are clear. 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.[13]

Here Sumner relies on a list of processes that, according to Sumner, seem *not* to produce autonomous preferences. And, frankly, the list seems about right (whether or not there is an underlying theoretical unity to such processes). The processes noted by sumner seem clearly to interrupt the extent to which a genuine preference can rightly be described as expressing *my* values.

Take now the time-slice accounts. TA1 also seems plausible strictly as a theory of autonomous preferences. I might maintain a particular preference for something rather than another thing. But one thing that might indicate the extent to which this preference is autonomous is my *attitude* toward this very preference itself. For instance, imagine that I am a drug addict.[14] I might prefer a dose of the drug to which I'm addicted rather than refraining from taking that dose. But is that preference autonomous? We won't know the answer to this question unless we know whether I endorse that preference, or take a pro-attitude toward it. (What *sort* of pro-attitude is proper for such endorsement is left unaddressed here, but, classically interpreted, it is a form of second-order desire or second-order preference, a "preference to prefer" the object of one's first-order preferences.) If I'm perfectly OK with my preference, despite the fact that I'm addicted, one might say that it is, in fact, autonomous: I endorse this preference (say, I "prefer to prefer it", or "desire to desire" taking the dose). Indeed, this account of the nature of autonomy forms the backbone of Frankfurt's influential account of the autonomous *will*, along with a number of influential accounts of the nature of personal value.[15]

TA2 adapts a thought common to historical accounts specifically for time-slice accounts. As noted above, many believe that autonomous action is taken on the basis of reasons, at least reasons the person in question recognizes. But we might say the very same thing about autonomous preferences: when I *maintain* a particular preference for Diet Coke rather than Diet Pepsi *on the basis of reasons*, say, because the former tastes better, or because it contains fewer harmful chemicals, or because I like the color of the can, or because I find the advertisements less annoying, then we can

[13]L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Oxford University Press, 1996), 171.

[14]Cf. David Lewis, "Dispositional Theories of Value" in *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000), 70-71.

[15]Harry Frankfurt, "Freedom of the Will and the Concept of a Person" in *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988); Lewis, op. cit.; Peter Railton, "Facts and Values" in *Facts, Values, and Norms* (Cambridge: Cambridge University Press, 2004).

say that the preference in question is autonomous. It is reflective of genuine values that I maintain.

TA3 reflects the generally intuitive thought that the nature of autonomy is closely connected to the nature of control. When I act autonomously, for instance, this seems to entail that my action was *self-authored*, rather than controlled by external forces. But we might put the same thought to work when it comes to autonomous preferences. If I have control over a preference, whether or not to maintain it or refuse to do so, it could be that when I maintain it, this preference is autonomous, just as an action or choice is autonomous to the extent that I have *control* over whether I perform the action in question or not.

Finally, TA4 takes a slightly different tack. A preference is autonomous, on this view, not to the extent that I endorse that preference, but to the extent that I endorse the *object* of that preference. One might think that this account is relatively thin, insofar as to prefer something just *is* to endorse that thing. But TA4 requires an additional condition: that one would endorse the object of the preference in question under a suitably specified set of idealized cognitive conditions. So, to take a simple case, it could be that I prefer some particular object just because I fail to maintain sufficient information about it.[16] But I would not prefer that object, would not endorse the object of my preference, were I to maintain such information. In this case (depending on how one understands the nature of the idealized cognitive conditions in question), the preference is not autonomous; the object of this preference would not be endorsed under idealized cognitive conditions.

It seems to me plausible, then, to say that each of these accounts connects with a strand of thinking about the nature of autonomy or the nature of autonomous preferences, specifically. Of course, insofar as each of these accounts focuses on a slightly different strand, they will not all be compatible. But the task of this paper is not to adjudicate between rival conceptions of autonomous preferences, but rather to determine whether *any* reasonable competitor accounts can adequately account for the normative authority of preferences, insofar as adaptation is a threat to such normative authority. I begin this investigation in the next section. One short note before I begin: this set of accounts of the autonomy of preferences is certainly not exhaustive. But, or so I shall argue in the conclusion, taking these accounts together allows us to draw an important conclusion about the relationship between adaptive preferences and *any* possible account of the autonomy of

[16]Daniel Haybron, *The Pursuit of Unhappiness* (Oxford: Oxford University Press, 2009), 185; Sumner, 139.

preferences.

4. HA1

According to HA1, autonomous preferences are developed by processes that are engaged in by people on the basis of reasons. But this cannot be the correct account of the normative authority of preferences. To see why, one need only note that adaptation to one's circumstances is a phenomenon that can occur for *perfectly good reasons*. Anyone, for instance, would see a reason to be more cautious around manholes if, in fact, one has recently fallen into one. But it also seems plausible to say that the mechanism of preferential adaptation can occur for similar reasons. When our preferences are frustrated, this is painful, it makes our lives worse than they might otherwise have been. And if there is no hope of fulfillment of those preferences, or if those preferences seem unlikely to be fulfilled, it may seem not just natural but positively *rational* to adapt our preferences to our circumstances.

But the rationality of the phenomenon of adaptation, when in fact it is rational, does nothing to vindicate the normative authority of preferences that undergo such processes of adaptation. To see why, consider the potential influence of preferences on theories of social choice. Note that a significant way in which preferences are adaptive are on the basis of existing social realities such as oppression, poverty, lack of opportunity. But adaptation to these realities can be rational for all the reasons just mentioned. But if and when policymakers are choosing to change social structures, or to assess the quality of such structures, it would be "ethically deeply mistaken"[17] to assess their quality in light of our rational preference revision to the social status quo.

I hasten to note that this problem also plagues a further account of the nature of autonomous preferences, viz., those preferences that are not formed as a result of *covert influence*.[18] Covert influence is surely a feature of some instances of adaptation, but is not necessary. An individual's adaptation to his or her circumstances, like my adaptation to Columbia rather than Harvard, can be perfectly up-front. Indeed, I might choose, for perfectly good reasons, to adapt in this very way. But this doesn't mean that there isn't anything "wrong" with adaptive preferences, as Colburn claims. Their normative authority remains suspect even in light of the fact that they chose to develop them for perfectly up-front reasons. (Much of Colburn's discussion

[17] Amartya Sen, *On Ethics and Economics* (Oxford: Blackwell, 1987), 46.

[18] Ben Colburn, "Autonomy and Adaptive Preferences" in *Utilitas* 23 (2011), esp. 64-70.

is motivated to distinguish adaptive preferences from "deliberate character planning,"[19] which is the "intentional shaping of desires" to, more or less, what one can get.[20] But if we're concerned about the normative authority of preference, this is a distinction without a difference. As I understand it, adaptive preferences are characterized as *adaptations*; this adaptation can be conscious or unconscious, can take place "behind one's back", or as a result of deliberate choice. But this does nothing to alter the plausible judgment that such adaptations do not reflect *the real me*.)

Furthermore, however, there is good reason to believe that some normatively authoritative preferences will be the result of processes that I *do not* engage in on the basis of reasons. For instance, the fact that I grew up in certain formative years around fans of the estimable NFL franchise The Washington Redskins was a process that led me to prefer—even very deeply prefer—that the Redskins win, rather than that they lose. But it would be extremely implausible to say that this process is one that I undertook or engaged on the basis of reasons. This process, rather, *just happened to me*. But this preference is not adaptive and is surely normatively authoritative. If, for instance, the owner of the team wished to be genuinely beneficent to me, one way to do so would be to ensure that the team won more often, etc.[21]

If this is correct, it cannot be right to say that normatively authoritative preferences are autonomous *if* the right account of autonomous preferences is HA1. A process of adaptation can be engaged in on the basis of reasons, and indeed can be straightforwardly rational, even if those preferences clearly lack normative authority. In addition, straightforwardly normatively authoritative preferences can be the product of processes that were not engaged in on the basis of reasons.

5. HA2

According to HA2, autonomous preferences develop according to a process that the person whose preferences they are would endorse, perhaps after rounds of cognitive, idealized reflection.

HA2 can solve one of the problems that faces HA1. For instance, it could be that in considering the process that led to my preference that the Washington Redskins win, I am perfectly willing to endorse it. This process

[19] Colburn, 55; See also Elster, 224.

[20] Elster, 224.

[21] This is a problem for Colburn's account, as well. Cf. Bruckner, "Colburn on Covert Influences" in *Utilitas* 23 (2011), 455-6.

may well conform to my general values or, at the very least, would not alienate me to any meaningful degree. But notice that HA2 retains the first problem with HA1. Take some process P, the end result of which is an adapted preference. Given the state of the world, or the state of my society, or some particular fact about me, it may be very clear that I cannot satisfy some preference of mine, or it may be good for me to develop some other preference more in line with the status quo. And if this is the case, I will have reason to engage P. But insofar as I have reason to engage P, it seems difficult to understand why I would regard P as a process unworthy of endorsement. After all, engaging it is rational given the state of the world, my society, or myself.

One might revise HA1 and HA2 in light of the problems noted here. One possibility is to say that an autonomous preference is one produced by a process engaged in—or endorsed—only for the *right kind of reasons*.[22] Note the shape of the reasons that lead me to adapt my preferences: such reasons are instrumental or strategic; they are not given my assessment of the *per se* or *intrinsic* value of the processes involved. Perhaps these are not reasons of the right kind. But what are? To salvage HA1 or HA2, any such account would have to rule out the possibility that the reason in question is strategic or instrumental in the way the cases of revision I've so far explored are. But it would have to do this without ruling out the normative authority of preferences we generally recognize as authoritative. But this is an extremely difficult problem. For instance, take one suggestion. One might say that the wrong kind of reason to endorse a process, or to engage in a process, is a purely strategic reason, i.e., a reason based simply on the instrumental effects of a particular process. This proposal cuts too deeply. Take a freshman entering college who has to decide whether or not to "go Greek", i.e., join a fraternity or sorority. Imagine that this person is neutral either way regarding the intrinsic benefits of going Greek, but nevertheless has some realization of the instrumental benefits of Greek-dom, viz., increased social connections, a leg-up in campus political races, etc. This person goes Greek for these reasons, and over time develops a strong preference for the Greek organization to which he or she belongs. In this case, the *process* that resulted in the preference was embarked upon for almost exclusively strategic or instrumental reasons. But we wouldn't say that the resulting preference for the Greek organization to which this person belongs lacks normative authority. In addition, it seems likely (or at

[22]For a discussion of the issue I note here, see David Sobel, "Full Information Theories of Well-Being" in *Ethics* 104 (1994), 793n19.

least we can imagine that it is the case) that this person doesn't *endorse* the process by which he or she develops the preference in question (i.e., Rush week, or whatever other "loyalty building" exercises one engages in) for other than instrumental reasons. But the preference for his or her Greek house nevertheless maintains normative authority, and should be treated as such.[23]

But let's leave aside this point for the moment. There remains an important problem with HA1 and HA2 even if we are able to offer a plausible account of the "right kind" of reasons. It is this: to recognize a reason is to take some sort of positive attitude toward a state of affairs, fact, or other entity. But there is no guarantee that *these positive attitudes* themselves are not the product of adaptation. I may very well recognize a *non-strategic* "right kind" of reason to, say, embark upon a process of adapting my preferences to my condition. But *this recognition itself* could be the product of problematic forms of adaptation; I might adaptively see a reason to remain in the midst of, rather than escape, poor or oppressive conditions and thus come to, non-strategically or for whatever "right kind" of reason, embark upon or endorse a particular preference-formation process *that further adapts my preferences to such conditions.* And thus insofar as the recognition of reasons itself is susceptible to the same failures of autonomy, we cannot look strictly to one's assessment—even "right kind of reason"-based assessment—of preference-formation processes for the sake of adequately ruling out adaptive preferences.

6. HA3[24]

HA3 is different. Rather than focusing on one's *endorsement* of, or *decision to undertake* a particular preference-formation process, it focuses on the content of the process itself, whether endorsed or not. Take the prototypical example of an insidious adaptive preference, i.e., a preference that one happens to have as the result of *brainwashing.* In this case, the problem with this preference seems to be the method by which it was developed, and the fact that this method is entirely antithetical to the autonomy or reasoning capacities of the agent in question. In other words, the process by which the preference was instilled was not one that instills a preference in a person *on the basis of reasons.* Rather, it instills a preference on the

[23]For a further discussion of the problems noted here, see Dale Dorsey, "Subjectivism without Desire" in *The Philosophical Review* 121 (2012), 412-415.

[24]I'd like to thank Antti Kauppinen for excellent and thoughtful comments and conversation about this view.

basis of, well, brainwashing. The same might be said for other examples of adaptation that are generally regarded as non-normatively authoritative. If I develop some preference simply as a result of oppression, I don't develop it, or so one can assume, on the basis of reasons.

But this account soon crumbles. Leaving aside the fact that it seems to succumb to the very same problems that characterize HA1 and HA2 (for instance, a process of adaptation *itself* could be a process that generates new preferences on the basis of reasons—even "right kind" reasons—given that the recognition of reasons can be colored by the process of adaptation), this account is under-inclusive. It is surely not the case that I developed my preference for Washington Redskins victories on the basis of anything like a *reason*; it just simply developed as a result of my social circumstances. No reasoning or engagement of my rational capacities were involved. This doesn't mean, however, that such a preference is not normatively authoritative.

Indeed, it would seem that *many* of the most hum-drum preferences we maintain are not the result of processes that instill preferences on the basis of reasons. My preference for black coffee over coffee with cream, for instance, was not the product of a reason-based process, but was developed, presumably, simply given the fact that my first experiences with coffee were without cream. My preference for Beethoven rather than Mozart might be a result of music my parents played in the house when I was an infant. None of this is plausibly regarded as the development of a preference as the result of a reason-based process. And hence, it seems to me, even if HA3 were not over-inclusive, it is certainly under-inclusive and hence must be rejected.

7. HA4, and a General Argument Against Historical Accounts

The final historical account eschews the possibility of finding any unifying or underlying feature of the "right" processes by which normatively authoritative preferences develop. This account instead simply settles for an "objective list", or a list of processes that can plausibly be said to generate autonomous preferences, and, thereby, a list of processes that can plausibly be said to generate non-autonomous preferences. Brainwashing will go on the latter list. Simple processes of preference formation while a child will, for instance, go on the former list. And so on.

One could critique HA4 for refusing to offer any further account of the rationale for inclusion of any particular process on the list of right or wrong

processes.[25] But this would be to miss the point of the proposal. The proposal on offer is that there is no such rationale. And hence to complain on this basis would be to pound the table in favor of a more unifying account, an account HA4 denies the existence of. One may find it implausible to believe that there should be no more underlying unifier. But, as we have so far seen, it is difficult to find such a unifier that could plausibly explain the relative normative authority of preferences instilled by some processes rather than others. And so if one is committed to an historical account of autonomous preferences, then an historical account would seem to require abandonment of the hope for such a unifier.

But the problems with this view, I think, are shared by all historical accounts. First, it seems right to say that any particular preference formation process can produce preferences that are or are not problematically adaptive depending on *other* facts about the person in question. For instance, consider the example of preferences formed based on one's early social community, such as my preference for a Redskins touchdown. It seems right to say that *in some cases* this very process can also yield preferences that are problematically adaptive and hence non-normatively authoritative. For instance, consider the possibility that one's social condition is oppression- or poverty-ridden. If the individuals in my community, whose preferences are already adaptive, instill in me similarly adaptive preferences (say, preferences against social advancement, or preferences to remain in my poverty-stricken condition), few would say that this preference is normatively authoritative, or should be taken seriously in an assessment of my own good or of the success of social policies. And if this is correct, it would seem that a particular process of preference-development or installation is neutral with regard to the extent to which that preference is normatively authoritative. It depends on the person and the preference.

But what is the explanation of this? I think the most plausible one is that in the case of the Redskins, there is no particular conflict between the preferences I have and any other attitudes I have. On reflection, I would *judge* a state of affairs in which the Redskins won a good one, or at least one to which I have an unproblematic attachment. But in the case of a preference for one's own poverty, it seems plausible to say *at first glance* that anyone considering, in the cold light of day, their own conditions as poor is likely to admit that the preference they maintain is maintained simply as a strategic device, or that given full information about one's self

[25] For a similar critique of the objective list theory of welfare, see David Brink, op. cit., 32.

and one's circumstances, he or she would judge that that which one prefers is unattractive or lacks value. Given full information about the state of the world, it seems plausible to say, any person who maintains a preference for a life of poverty would be likely to judge that life undesirable in comparison to other lives. And hence the problem of adaptive preferences is not trying to ferret out the processes by which these preferences were formed, but rather by coming to a kind of consistency or coherence among one's evaluative attitudes.

Here's another way to put this critique. The normative failure of adaptive preferences seems traceable to the fact that such preferences, broadly speaking, do not represent my genuine values; in this sense they are not autonomous. This helps to explain why my preference for Redskins victories is authoritative when my preference for conditions of squalor and poverty are not. But the extent to which a preference is reflective of *my values* seems more-or-less neutral with respect to the *processes* by which it developed.[26] Whether a preference is reflective of my genuine values has to do with that preference's connection to that person's wider evaluative attitudes. Thus even if we offer an account of the relevant preference-formation processes that eliminate all and only adaptive preferences, this account will not offer an account of autonomy that we have been seeking, viz., an *explanation* of our normative distrust in adaptive preferences. The processes by which adaptive preferences are developed seem to me evaluatively epiphenomenal. Instead, the right explanation is to be found in the extent to which a preference does, or does not, capture my genuine values. And this is not an historical, but rather a *time-slice*, property.

8. TA1-TA3

Time-slice accounts differ from historical accounts insofar as they identify autonomous preferences not on the basis of how these preferences came to be or the processes by which they were developed or instilled, but rather on the basis of facts about the relation between that preference and other psychological attitudes (perhaps counterfactual or idealized) of the person in question.

But this doesn't mean that time-slice accounts avoid the problems that plague historical accounts. Indeed, the first three seem to suffer from the same general problem that plagued the first three historical accounts. For

[26]For a more substantial argument on this point, see Donald Bruckner's paper, "In Defense of Adaptive Preferences" in *Philosophical Studies* 142 (2009).

instance, take TA1. TA1 holds that a particular preference is autonomous to the extent that it is endorsed, perhaps under idealized reflection. In this way, TA1 is similar to certain theories of the nature of the good. For instance, Peter Railton writes: "Let us then say that an individual's *intrinsic* good consists in attainment of what he would in idealized circumstances want to want for its own sake... were he to assume the place of his actual self."[27] David Lewis says something similar: "[A person] does not value what he desires, but rather he values what he desires to desire."[28] Here it would appear that a particular preference or desire is *evaluatively* authoritative (i.e., its object is intrinsically valuable) to the extent that this desire is itself endorsed: someone *wants to want* something, or *desires to desire* it. Of course, a mere second-order desire isn't the only way one might construe the right sort of endorsement. Bruckner suggests that the relevant endorsement of a preference is an "all-in judgment that can conflict with a second-order preference."[29]

But this proposal gets into precisely the same trouble as HA2. Though it is possible that my idealized self is perfectly happy to endorse my preference for Redskins touchdowns, it is also very likely to be the case that this idealized self will endorse perfectly rational adaptive preferences given the various reasons one might have to develop them.[30] This is clear in, e.g., Bruckner's proposal. According to Bruckner, a gymnast who endorses her adaptive preference to compete in regional tournaments rather than the Olympics, given her lack of ability to do so, would maintain a normatively authoritative preference. But this proposal is implausible. Though it may be rational[31] for her to maintain this preference, such a preference lacks

[27]Railton, 54-5.

[28]David Lewis, "Dispositional Theories of Value" in *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press, 2000), 70-1.

[29]Bruckner, 317.

[30]One might try to revise TA1 in light of the problems noted here, and again identify a set of "right reasons" to endorse a particular preference. For instance, one might say that autonomous preferences are those that are endorsed *for their own sake*. But this proposal fails. It seems odd to say that my preference is autonomous only if I desire or endorse that preference *for its own sake*. I don't endorse a preference to be a philosophy instructor for its own sake; I endorse that preference because *I value being a philosophy instructor*, and hence having that preference is conducive to that which I value. But this doesn't mean that such a preference lacks normative authority or is problematically adaptive. There are other ways to construe this idea; regrettably I don't have the space to discuss them all, but I have done so elsewhere. I refer the reader to "Subjectivism without Desire", 413-415.

[31]Indeed, Bruckner sometimes suggests that his proposal is simply that endorsed preferences are "rational". If this is correct, then I agree. But the tricky aspect of adaptive

the normative roles we assign to normatively authoritative preferences. We should not believe, in other words, that performing in regional competitions is better for this person than performing in the Olympics *would be*, despite her endorsed adaptation.

TA2 holds that a preference is autonomous insofar as the person whose preference it is maintains it on the basis of reasons. But as I have so far been at pains to argue, there are *reasons* to adapt one's preferences, and hence saying that autonomous preferences are those that individuals maintain or possess on the basis of reasons isn't going to solve the problem of normative authority. Though, perhaps, it remains a reasonable representation of one facet of the *autonomy* of preferences.

TA3 takes a slightly different tack with no better results. One might understand the notion of autonomy as picking out a form of control someone might have: autonomous actions are actions that, generally speaking, one has control over. Thus autonomous preferences are those over which one has control. If I can choose to maintain a preference or to rid myself of it, and I choose to maintain it, it would appear that this preference displays a form of autonomy in a perfectly respectable sense of that term.

Looking closely at this proposal, it is obvious that it cannot work. This is because the reasons to maintain adaptive preferences remain the same whether one has control over them or not. And so *if* one has control over whether to adapt to one's preferences or not, there will be a number of cases that render adaptation rationally justified. So far, time-slice accounts fare no better—and perhaps fare even worse—than historical accounts.

9. TA4

TA4 takes a substantially different approach than the previous accounts. This account states that autonomous preferences are for objects or states that are themselves endorsed or valued *under conditions* of cognitive idealization. Here's a reason to think that this might work. One plausible thought is that an individual's *genuine* values are those that he or she *would* develop if only the right sort of cognitive conditions held. Take, for instance, my adaptive preference for conditions of poverty or oppression. Though, as I actually am, I endorse this state, it is hard to see how I could or would endorse or value that state if I adequately appreciated the status of alternative ways

preferences is that their rationality does not entail normative authority. But if Bruckner wishes to claim that such preferences "ought to play the same role in our rational deliberation as the rest of our preferences," this seems incorrect for many cases he cites, for reasons discussed here. (Bruckner, 311.)

of living. If I fully understood the distinction between a life of poverty and a life free of destitution, and this distinction was made clear and vivid to me, it would be unlikely that I would keep such a preference.[32] Furthermore, it seems plausible that simple information about the world would not alter my general preference that the Redskins win the Super Bowl, as compared to some other football team, etc.

Notice that the "cognitive idealization" clause is essential for the proposal in question. Indeed, even a bare preference is itself a form of endorsement of a particular object. If I prefer to attend Columbia to attending Harvard, this is a form of endorsement of my attendance at Columbia. What is essential for this account, however, is the conjecture that cognitive idealization will slough off those preferences one has that are *merely* adaptive or do not express the genuine values such preferences are adaptations from.

Of course, this is a *big conjecture*. Is there any program of cognitive idealization that would guarantee that all and only non-adaptive preferences would be the result? I think a plausible answer begins by noting that a person's genuine values, as opposed to simply adaptive ones, must avoid being those that are *simply* a consequence of the world as it is. Elsewhere, I have argued that we can offer an account of an individual's preferences that is independent in this way if we combine a *full experience* constraint with a *coherence* constraint.[33] My account of this view's virtues will be *very* sketchy, but take, for instance, my preference to go to Columbia rather than Harvard. This preference clearly depends on the contingent fact that I didn't get into Harvard. So what would my preference be if I examined the virtues of going to Harvard rather than Columbia *independently* of this fact? What if I fully understood and experienced both college careers, and revised my preferences in light of this experience toward a goal of coherence? Clearly, my adaptive preference would be revised away: this is because in considering my options independently of my *actual* life, the "deep" values I maintain that point toward Harvard rather than Columbia will resurface, and (given their depth) will be maintained in any revision toward coherence.[34] The

[32]See, for instance, Richard Brandt, *A Theory of the Good and the Right* (Oxford: Oxford University Press, 1978), 113-126.

[33]Dorsey, "Preferences, Welfare, and the Status-Quo Bias" in *Australasian Journal of Philosophy* 88 (2010), §4.

[34]Much of this explanation relies on what it means to render a set of preferences coherent. See "Preferences, Welfare, and the Status-Quo Bias", 547-549. In essence, the general thought is that in cases of recalcitrant preferences, one revises based on a broadly Quinean rubric of "minimal mutilation", keeping fixed the "deepest" or most firmly held preferences, and/or those that would require more, rather than less, overall revision to one's preferential set.

same seems correct about, say, adaptive preferences toward one's poor or oppressed conditions. As I come to possess full and adequate information about the various alternative styles of life I will develop potentially very deep preferences (preferences for greater freedom and opportunity, say) that are incoherent any adaptive preference toward poverty. If I have full experience of the alternative (even inaccessible) ways I might live, and my preferences are broadly coherent with what I find most important *given* such experience, we can and should say that the preferences I maintain are not dependent upon any *particular* style of life or any particular facts about the world as it is. And hence such preferences cannot merely be adaptations to the way of life I maintain or the world around me. Rather, such preferences are reflective of what I *genuinely* value.

Much more needs to be said to defend this claim and so I won't put much weight on it here. However, even if there is a method by which to avoid cases of adaptation we've so far been discussing, there remains a problem with TA4. Let's call this the problem of "deep adaptation".[35] No matter how much cognitive idealization one undertakes, if this form of cognitive idealization is really going to be rooted *in the person in question*, then it would appear that the only thing required for any rubric of cognitive idealization to retain adaptive preferences is for those preferences to be central to one's psychology. If my adaptive preference becomes entrenched in my own psychology and self-identity, it seems hard to see how any preferences I might develop as a result of increased information or experience would themselves be strong enough or deep enough to override this deeply held adaptive preference in any bid for coherence.[36] But *however* the proper form of cognitive idealization is understood, there is no guarantee that adaptive preferences won't end up as some of the *deepest*—and most central—preferences one has.[37]

For instance, imagine that at some point in my life I come down with a chronic illness, say, type 1 diabetes. One method of coping with the disappointment of this illness is to strategically revise my preferences and evaluative attitudes: I come to hold that being diabetic is "central to my identity", say, and that monitoring my blood sugar manually is a form of "being in touch" with my body's chemical processes, etc., etc. This preference is adaptive. But the more I come to identify with it, the deeper it becomes, and the less likely it becomes that any plausible account of my gen-

[35]Dorsey, *The Basic Minimum: A Welfarist Approach*, 100-108.

[36]Dorsey, "Preferences, Welfare, and the Status-Quo Bias", 551-2.

[37]For a further argument to this effect, see Jennifer Hawkins, "Welfare, Autonomy, and the Horizon Problem" in *Utilitas* 20 (2008), 165-167.

uine values will *not* include it. This result generalizes. Whatever program of cognitive idealization one prefers, it is possible, given the plasticity of human psychology, for an adaptive preference to pass the relevant idealization tests, and hence (according to TA4) remain normatively authoritative.

10. Conclusion

So what has been shown? So far no account of autonomous preferences has been able to adequately expunge adaptive preferences without also expunging perfectly respectable ones. But what conclusion are we to draw from this? Where are we to go from here? One possibility would be to offer some further account of the autonomy of preferences. But reflection on the failure of the accounts discussed here seems to offer a more general conclusion. As it happens, any plausible understanding of a person's true or genuine preferences or values is necessarily subject to problems of adaptation. Sometimes adaptive preferences, perhaps simply as a result of time, will form a crucial *part* of an individual's deepest self-understanding or point of view. (This was clear in the failure of TA4.) In short, adaptive preferences will, in many cases, form a significant aspect of our individual, autonomous, points of view. If so—necessarily—no adequate account of the autonomy of preferences will expunge all instances of adaptation.

In light of this, one might be tempted to shift the focus from the autonomy of a given preference, in determining its normative authority, to some other feature. In other words, if treating *autonomous* preferences as normatively authoritative cannot prevent adaptive preferences from possessing normative authority, we must focus on some other property that preferences possess or do not to explain the normative failure of adaptive preferences.

However, this strategy seems to me a poor fit for any view that seeks to take preferences seriously. In rejecting the normative authority of perfectly autonomous preferences, we lose a central motivation—explored above—for granting preferences a normative role. It seems plausible to say that preferences are important to take seriously, in coming to accounts of the nature of a person's good, or accounts of prudential or beneficent reasons, or accounts of proper social choice, because a person's preferences are supposed to represent something a person *values*. They are supposed to represent that person's way of evaluating the world around them, which itself is a significant normative fact. But if we reject the authority of preferences that, in fact, form a significant part of the way a person sees and evaluates the world around them, it's unclear to me why we should trust preferences to be normatively authoritative at all. Of course, if an individual's preferences

are sadistic or shallow, or display various other *external* problems, we might downgrade them for this reason. But it's hard to see why, if a preference adequately corresponds to the various external measures (including morality, respect, objective value, etc.) *and* adequately corresponds to the relevant *internal* measure (i.e., the expression of genuine values), we should fail to grant it normative authority. That is, unless we refuse to accept the claim that motivated a normative interest in preferences in the first place.

I think there are two ways of addressing our current predicament. Call the first the "pessimistic" conclusion. This conclusion is motivated by the fact that to reject all adaptive preferences is to substantially weaken the motivation for taking preferences normatively seriously. The result of this conclusion is that preferences simply can no longer be trusted to play the normative role(s) to which they have been tasked: no plausible representation of *my* preferences can rule out all instances of adaptation. Perhaps the good, moral obligation, and social choice (or any other normative domain, for that matter) should look beyond my own values, no matter how genuine they are.[38]

For my money, the pessimistic conclusion would be a hard pill to swallow. Imagine, for instance, how strange it would be to deny that an individual's preferences have something to do with with intrinsic value. It could be, for instance, that I strongly prefer a career as a professional baseball player to a career as a professional football player, just for its own sake. But it would appear that to deny the relevance of preferences would be to say that this fact can have nothing to do with the relative value of being a professional baseball versus football player for me. Any such value must be the outcome of facts that are independent of my own preferences. This is not incoherent to say, of course, but it is very implausible. Surely the fact that I prefer being a baseball player, other things being equal, makes it the case that being a baseball player is better for me than being a football player. This is not to say that facts beyond a person's preferences aren't relevant to value. But preferences surely *are* relevant in this way. Something similar should be said about social choice. Although there are plenty of arguments that seek to show that social choice should be neutral when it comes to conceptions of the good,[39] surely if everything else is equal, making it the case that more rather than fewer of people's preferences are satisfied is something to be accepted in a social policy.

[38] This conclusion is urged by Hawkins, op. cit.; Nussbaum, op. cit.; Sen, op. cit.

[39] John Rawls, *Political Liberalism* (New York, NY: Columbia University Press, 1993); Jonathan Quong, *Liberalism Without Perfection* (Oxford: Oxford University Press, 2010).

But the pessimistic conclusion is not the only option. Why believe that all adaptive preferences must lack normative authority? Especially if we believe that *problematic* forms of adaptation are those that "mask" an individual's genuine conception of the good, why must we say that adaptive preferences that *do not* mask such conceptions of the good fail to maintain normative authority? After all, "deep" adaptive preferences surely capture, rather than obfuscate, what I genuinely value—they plausibly represent *my* conception of the good. And so the "optimistic conclusion" is this: there is, in fact, a reasonable method by which to account for the distinction between normatively authoritative and non-normatively authoritative preferences. Assuming a proper round of cognitive idealization can be found—which I admit is a rather big assumption—we might say that it is TA4 (or something like it).[40] But this entails that "deep" adaptive preferences are normatively authoritative.

I think this is the right answer. But I will not argue for it here. It suffices to note that there is an important choice to be made in our treatment of the normative authority of preferences, especially adaptive preferences. Either we reject the normative authority of all adaptive preferences—and with it reject the motivation for granting normative authority to preferences at all—or we allow that some adaptive preferences, those that are suitably "deep", are normatively authoritative. Which choice to make is something I leave, at least for now, open.

[40]Indeed, it seems to me that this is precisely the correct answer, and I have developed it in detail elsewhere. See Dorsey, *The Basic Minimum*, ch. 3.