Authenticity and the Third-Person Perspective

Chapter 7 in G. Levey ed. *Authenticity, Autonomy and Multiculturalism* (New York: Routledge, 2015): pp. 121-142.

Ben Colburn*

In this chapter, I make two proposals. First, "authenticity" is a label for the property possessed by all and only those preferences whose satisfaction contributes to our lives going well. Second, the property in question has to do with those preferences' explanations: our preferences are authentic just in case they do not have covert explanations, which is to say when the true third-personal explanation of our preferences is necessarily hidden from our first-person perspective.

This theory is perhaps surprisingly simple, in light of the variegated usage of the concept of authenticity in contemporary moral and political philosophy. In what follows, I deliberately avoid beginning by engaging directly with much of that discussion, mostly because it seems to me to have become so very muddled that close analysis of the literature is unlikely to yield anything useful. Some have taken this muddle to indicate that the concept of authenticity itself is ill formed, and that we would do better to abandon it in favor of something clearer, for example Lionel Trilling's famous exhortation to abandon authenticity in favor of the earlier virtue of moral sincerity (Trilling 1974), and Veit Bader in this volume (Chapter 6). What follows here is an attempt to avoid that eventuality. For those dissatisfied by my setting aside existing theories of authenticity, I address at the end of my first section how my account of authenticity relates to them.

Following in the footsteps of Edward Craig (1990) and Bernard Williams (2004) in their work on knowledge and truth respectively, I begin by offering a constructive genealogy for a theory of authenticity. Asking why we need such a theory in the first place allows us to identify a general *concept* of authenticity, functionally defined by the role that it plays in our moral thinking. That the concept is functionally defined allows us moreover to derive a set of desiderata against which particular *conceptions* of authenticity can be judged. Having established this, I go on, in my second part, to discuss various attempts to formulate such conceptions, and show why they fail. Drawing on the lessons of those failures, in my third part, I set out my own conception, and show it meets the desiderata that its rivals don't. Finally, I explain why my conception of authenticity seems uniquely well placed to meet those desiderata, and hence why this seems to me the best chance for a defensible theory of authenticity.

A Brief Genealogy

The concept of authenticity arises when we seek to answer the following question: What is the relationship between the satisfaction of an individual's preferences and how well her life goes as a whole? In what follows, I set out a view on which authenticity should be understood as a component needed for any view that seeks to avoid two unattractive answers to this question.

The first such view says: there is no essential connection between preference satisfaction and well-being, because what makes a life go well is settled independently of the individual's preferences. For example, one might think that living well means striving, so far as possible, to live by the Ten Commandments. Insofar as one's preferences aim at something else, so much the worse for them; living well requires either changing or ignoring them.

The second unattractive view says: there is nothing more to well-being than the satisfaction of an individual's manifest preferences. Over the course of a life an individual will have lots of preferences, and how well her life goes is just a function of how many of them are satisfied. We might have to do some work to determine precisely what that function *is*, of course, but in general all and only the satisfaction of manifest preferences is what counts.

Many philosophers – especially those who are sympathetic to some ideal of individual autonomy as a component of well-being – find neither position attractive as they stand.[1] Reflecting on the reasons for that will explain where our need for a concept of authenticity arises, and hence allow us to see what would count as a successful conception of authenticity. I should say that what follows is not an attempt to give conclusive arguments against either position. Rather, I seek to rehearse some of the reasons which might motivate one to reject them, and then to reflect on what has to be true if one *is* going to reject them.

Views in the first camp – Objective List theories, as Derek Parfit dubs them (1984: 493–502) – face a number of awkward questions. For one thing, their proponent must be able to explain how we should identify what sits on the objective list, and whence it would derive its normative force. For another thing, because what sits on the objective list is not determined in any way by an individual's own mental states, the theory raises the unpalatable prospect that one's well-being can be improved by things which one neither notices nor cares about. Now, reporting these gut reactions does not constitute an argument against Objective List theories, although as a matter of fact I do think they provide the bases of arguments that give us good reason to reject such theories. My mentioning them here, as I've said, has just the modest aim of identifying one of the central motivations behind the search for a theory of authenticity, which is that there is something wrong-headed about theories of well-being which are so detached from the individual's own perspective. So, my claim is conditional: to the extent that one shares these qualms about the Objective List theories, one should think

---

***** My thanks, for helpful comments and criticism, to David Bain, Daniel Elstein, Lindsay Hamilton, James Humphreys, Ian Law, Geoffrey Levey, Hallvard Lillehammer, Neil McDonnell, Michael O'Donnell, Neil Sinclair, Christopher Snow, Iona Stevenson, Alan Wilson, Alastair Wilson, and other members of audiences in Birmingham, Glasgow, Newport and Nottingham.

[1] This includes philosophers who take themselves to be defending some version of either view, since to make those positions defensible they modify them from the simple versions described here. For a few prominent examples, see Parfit 1984, Arneson 1999, Sumner 1999, Feldman 2004, Crisp 2006, and various papers in Olsaretti 2006.

that someone's own perspective (or, specifically, their preferences) play a central role in determining what counts as their life going well.

The preceding line of thought might incline us towards the second view essayed above, namely that what matters is just the satisfaction of our manifest preferences. However, that too is problematic. To start with, there are some preferences which are repudiated by the agent concerned. For example, imagine a reluctant smoker, who wishes she weren't addicted to cigarettes. Given that even she herself thinks that smoking makes her life go worse, it would be odd to say that her life goes better if she satisfies her desire for nicotine-laden smoke. Of course, we might just say that we should exclude from consideration preferences which are explicitly repudiated like this. But even if we grant that for the sake of argument, and therefore restrict ourselves to considering only non-repudiated preferences, the Manifest Preference Satisfaction theory has some surprising and unattractive consequences. People frequently have stupid, malevolent, inconsistent or changeable preferences. I might love making mud pies, or want to torture babies; I might desire both to eat chocolate cake and to eat only healthy food; I might vacillate perpetually between wanting to be an academic philosopher and wishing I could have been an architect instead. Taking those manifest preferences to determine what makes my life go well leads to a conception of well-being which is likewise stupid, evil, impossible to satisfy, or terribly unstable. For present purposes, all my argument requires is that at least some of those implications are likely to be seriously unintuitive.[2]

Furthermore, taking all manifest preferences at face value is implausible because the process of preference formation is itself a proper object of philosophical concern. For one thing, the institutions about which political philosophers argue – education, democracy, markets, advertising – have a profound effect on the content of people's manifest preferences. Indeed, that they do so is frequently part of their purpose, and always crucial to their effectiveness. There would be something peculiar about recognising that such institutions are susceptible to philosophical scrutiny and at the same time thinking that we should abstain from scrutinising the preferences that those institutions partially form. For another, many phenomena which we might worry about from the point of view of justice – entrenched sexism or racism, for example – are effective at least partially because they shape people's preferences. To illustrate: imagine a female worker who responds to endemic sexism in her workplace by downgrading her ambitions to reach the top of her organization. She has ceased to have a preference to become CEO of her company, now, and sincerely prefers a less powerful role. Even so, her current manifest preferences still merit a critical look. Even if we agree that we'd only now make things worse if we tried to interfere, we can still think there's something troubling about the relationship between her manifest preferences and how well her life goes. We can still ask whether her life mightn't go better if her preferences were otherwise, and whether she mightn't do better if her workplace had not conformed to

[2] In part three of my essay, I suggest that some of these counterintuitive views are bullets that we'll just have to bite once we've identified the best theory of authenticity, so I don't meant to imply that all of these views are obviously out of bounds. They do, however, provide us with some motivation to seek something else.

her preferences (which, remember, she came to have in response to its unfairness). These questions are open – but they are open only if we reject the view that how well people's lives go is simply a matter of how far their manifest preferences are satisfied. So, to the extent that we think such questions open, we should reject the Manifest Preference Satisfaction view.

As before, the preceding line of thought is not an argument against Manifest Preference Satisfaction theories, though once again I think it could provide the basis of such an argument. The point is just to reveal the second crucial motivation behind a theory of authenticity, namely thinking that the questions I've raised are open and substantive, rather than just being decided by the meanings of the terms "preference-satisfaction" and "well-being." If such questions are not settled analytically, then we should not think that our lives going well consists in just the satisfaction of all our manifest preferences.

What I have said so far offers some reasons to reject each of the Objective List and Manifest Desire Satisfaction theories. Those reasons motivate the two desiderata I reached in the preceding discussion, viz.:

1. Someone's own perspective or preferences should play a central role in determining what counts as his or her life going well; and
2. Not all manifest preference-satisfaction contributes to our well-being.

A theory of well-being which meets both these desiderata would be able to navigate a safe passage between the two unattractive theories listed above. What we need is a theory which takes an individual's preferences and perspective seriously, hence avoiding the Scylla of the Objective List theory, while retaining the resources for critical distance from what her manifest preferences are at a particular time, thereby escaping the Charybdis of the Manifest Preference theory.

To be successful, such a theory would need to give the following: a criterion that picks out a (local) property of preferences, identifying those whose satisfaction contributes to how well our life goes (considered globally). I suggest that the concept of authenticity identifies this function in our moral thinking. Particular *conceptions* of authenticity identify some particular property that purportedly plays this role. So, to illustrate with an implausible example: if one thought that well-being consisted in satisfying the preferences which one forms on a Wednesday, then one's conception of authenticity would identify it with the property of being-formed-on-a-Wednesday.

Setting up the discussion in this way allows us to make clear the levels of abstraction at which different philosophical disputes over authenticity sit. Some sit at the level of conceptions. Of those, some consist in the advancement of rival conceptions (a dispute between a Wednesday-preference theorist and a Thursday-preference theorist, for example). Others hinge on whether or not there is in fact a conception which fleshes out the concept of authenticity: one might, for example, argue that there is no such thing as authenticity because there is no property which plays the role identified above. Yet other disputes sit purely at the level of concepts. So, for example, someone might say that I have mis-defined authenticity, and that in fact the concept plays some different role in our moral thinking. Or, they might argue that there's no such thing as authenticity because there is in fact no clear and distinct role for it to play. This latter position ends up with the same conclusion as the

denial of authenticity at the level of conceptions, but the two reasons for reaching that conclusion are independent.

My main concern in this essay is with disputes at the level of conceptions: in the rest of this essay, I set out and defend my preferred conception of authenticity. Before I do, however, it will be worth responding to a criticism that might be levelled against my view on the concept of authenticity.

Above, I distinguished between two things: the global property of a life going well, and the local properties of agents at particular times which contribute to that global property. I use the term "authenticity" to refer to the second of those. In so doing it might seem that I exclude the possibility that authenticity might in fact be the goal of a whole life, and hence instead be more properly understood as the global property. That might seem perverse in light of the fact that it is treated that way in many familiar accounts of authenticity, for example the existentialist theories of Kierkegaard (1983) and Sartre (1948 and 1992), and the theory of the good life defended by Ronald Dworkin (2011).

That is a mistake, for two reasons. First, there is no such exclusion. If someone has a terminological preference for understanding authenticity as a global concept – whether that is of a whole person at a time, or of a whole life – then that will be consist with what follows; it will just be necessary to find some separate label for the local concept. Second, even the defender of a global concept of authenticity will need to identify and analyse the local concept I've labelled "authenticity." Remember, that concept just picks out whatever distinguishes those preferences whose satisfaction contributes to the global ideal (in my terminology, of well-being). Any proposed global concept of authenticity must say something about how we should regard individual preferences, and so it will need a criterion of my sort if it is to avoid being unattractive for the same reasons as the Objective List and Manifest Desire Satisfaction theories considered above. So, even if someone prefers to use the word "authenticity" to refer to a global ideal, what follows will be an important component of her theory.

Conceptions of Authenticity

A conception of authenticity will pick out a property (or set of properties) whose presence or absence distinguishes between those preferences whose satisfaction does or does not contribute to our lives going well. Put very loosely, it will divide the class of preferences into "good" and "bad." The distinction we are after is a familiar one. The difficulty comes when we try to find some coherent criteria which might underwrite our rough-and-ready judgments. In this part, I pick out some of those judgments, and then show how difficult it is to identify a general criterion, by running through some failed attempts.

Most preferences can be easily and intuitively sorted into one category or the other. For example: all other things being equal, it's good for me if my desire to drink orange juice is satisfied, and if I am successful in my ambition to climb Everest or cure cancer; but not if I satisfy a hypnotically implanted desire to behave like a chicken, or respond to harsh imprisonment by unconsciously downgrading my preferences to the level that they can be

satisfied by bread and water. Gerald Dworkin, in a slightly different context, suggests the following tasks:

> …distinguishing those ways of influencing people's reflective and critical faculties which subvert them from those which promote and improve them … distinguishing those influences such as hypnotic suggestion, manipulative coercive persuasion, subliminal influence, and so forth, and doing so in a non ad hoc fashion. (1988: 18).

Dworkin's remarks arise in the context of a discussion of autonomy, rather than authenticity: the comments specify what is needed for a theory of "procedural independence." Nevertheless, it's clear that procedural independence would indeed be a conception of authenticity on my taxonomy, since the former is a local property of preferences – to do with their formation – in virtue of which they can contribute to a global goal of autonomy (conceived of as a central component of well-being).[3] And the phenomena that Dworkin lists – hypnosis, coercion, subliminal influence – are the main intuitive examples that usually spring to mind as sources of inauthenticity.

Nevertheless, Dworkin goes on to concede that he has no general account of what procedural independence consists in: that is, no set of (non *ad hoc*) criteria distinguishing benign influences on preferences from the malign influences listed. His caution here is sensible. Faced with these intuitively plausible instances of formation mechanisms which lead to inauthentic preferences, various intuitively plausible criteria might spring to mind, but most of these lead immediately to problems. Some of them turn out to rule that *all* preferences are authentic, or all inauthentic; others are less strongly unintuitive but still pose problems insofar as they classify some paradigmatically problematic or unproblematic preferences the wrong way. In what follows, I discuss five examples with a view to motivating the thought that we need something different.

Authentic preferences are not externally caused

One possibility is that we might judge hypnotic suggestion, subliminal influence "and so forth" to be problematic because they are formed by a causal process outwith the control of the agent.[4] This plainly won't do. If the only authentic preferences are those in respect of

---

[3] I leave aside, for the moment, the question of the relationship between well-being and autonomy. In part three below, I show that a commitment to autonomy fits naturally with a concern for authenticity, and provides support for the particular conception of authenticity I seek to defend; but since I don't want to presuppose that the *only* motivation for a theory of authenticity is a commitment to autonomy, I defer discussion till then.

[4] On some interpretations Jon Elster draws the distinction this way when he contrasts the "purely causal process" of adaptive preference formation in which "the source of the preference change is not in the person" with the "conscious strategies of liberation" involved in conscious character planning: although he doesn't characterize it as such, the distinction he is trying to capture is the one between authentic and inauthentic preferences I am

whose formation we are able to slip the shackles of causation, the conception's being able to vindicate any preferences at all depends on our denying the metaphysical thesis of determinism. That is a serious theoretical cost in itself. It also makes our account of authenticity dependent on the wrong things. There are plenty of causally induced preferences which don't share the malign character of hypnosis or subliminal messaging. For example, it is at least partially through a causal process (to do with sound waves, neural pathways and so on) that my hearing the music of Handel induces in me a desire to hear more Handel in future. So relying on the causal/non-causal distinction won't do.

Authentic preferences are not caused by another agent

Maybe the problem is not so much causation in general, but causal influence specifically by another agent. If Jane hypnotizes me into wanting to give her all my money, then my preference in some sense "comes from her." Maybe that's what makes it inauthentic? Once again, I suggest not. First, almost any preference held by almost any agent is going to be causally influenced by some other person at some stage. So, construing authenticity as consisting in the absence of other agents' influence *tout court* would rule out almost all preferences. Second, when we look at the individual rulings that are made, the proposed criterion is at once inappropriately restrictive in some areas and inappropriately permissive in others. It rules out some other-influenced preferences which seem perfectly benign: it would say, for example, that satisfying my preference for listening to Handel doesn't contribute to my life going well if my preference arose because you lent me a CD to try and get me hooked on your favourite composer. And amongst the preferences it would vindicate are some which seem obviously malign: naturally acquired addictions, for example. In general, the proposal lionizes a class of preferences – namely those which lack other humans' agency in their formation – which don't seem particularly worthy of such treatment. The obvious conclusion to draw is that it is only *certain types* of influence by other agents that undermine the authenticity of preferences, but that leaves us with the same task we had in the first place, namely isolating the type of influence that is problematic.

Authentic preferences are consciously formed

A third possibility appeals to the distinction between unconscious and conscious preference formation. Maybe, we might think, the problem with the preferences listed by Dworkin is that they are not consciously formed. We might flesh out this proposal in two different ways, saying either that authentic preferences are those which we consciously create, or that they are those of whose formation we are conscious. As we will see below, I think there is a grain of truth in the latter of these possibilities. As they stand, however, each rules out too much.

---

pursuing here (Elster 1983: 109–110, 117). I am not convinced that this is the right way to read Elster, but the example illustrates an intuitive way we might try to draw the distinction. On the exegetical question, see Sandven 1995 (for) and Colburn 2011 (against).

Concerning the former, by ruling out preferences which we don't consciously create it threatens to rule out almost all preferences, for preferences induced through a conscious effort of will are in practice very rare. Nor do they seem to deserve any privileged status vis-à-vis our well-being. Does my successfully climbing Ben Nevis contribute to my life going well only if I sit down one day consciously to induce the relevant ambition, rather than acquiring it spontaneously on seeing the mountain for the first time? Surely not.

The latter proposal is problematic because there are lots of plausibly innocuous preferences of whose formation we are unconscious.[5] Suppose, for example, that I wake up one morning, find myself wanting to climb Ben Nevis, and can't work out why; maybe I saw a photograph of the mountain a few days ago, or recently read an article about the snows of Everest, but I can't put my finger on the exact cause. There isn't necessarily anything wrong about a preference whose genesis is unclear in this way, nor that my life is not made better by my satisfying it. So, insofar as we were trying to capture the intuitive distinction, the proposal that we understand authenticity to consist in conscious formation (either way) fails.

Authentic preferences belong to the true self

A fourth possibility is to draw the distinction at a deeper level. The problem with some preferences is that they're not truly *ours*: a hypnotically induced desire to give my money to cult is not really *my* preference, any more than a repudiated craving for a cigarette is. Other preferences bear a closer relationship to who we are: perhaps they constitute our true selves, or are the preferences which are held by our true selves.

This view is either empty or metaphysically dubious, depending on which of two ways we choose to disambiguate what is meant by the "true self."

If it is something that is constituted by our authentic preferences, then the proposal does not get us anywhere. What we were seeking was precisely a way of working out which preferences those are, and it doesn't help very much to say that they are all and only those preferences which make us who we really are, or something of that sort. It might be possible to construe that view so that the whiff of definitional circularity can be dispelled, but only if we have some other way of identifying the authentic preferences that are supposed to constitute our true selves, which is to say precisely the criterion for which we are currently searching.

On the other hand, we might think that the true self is prior to any of our preferences, and that the authentic preferences are those that bear a special relation to the true self. In advance of any such proposal being worked out in detail, one can't prove that no position of this sort could lead to a coherent, plausible and workable conception of authenticity. But the metaphysical, epistemic and meta-ethical costs of such a position are formidably high. Not

[5] I take this to mean "unconscious at the time of formation." Trivially, once we notice a preference, we will be conscious that it has been formed.

only would we need to have a defensible theory concerning what the true self *is*, to derive a workable conception of authenticity from it would need to be confident that it is the sort of thing we to which we have epistemic access, and has the normative status needed to vindicate preferences connected with it. This seems like such a tall order that it is would be worth looking hard for less costly alternatives before embarking on what will probably turn out to be a philosophical wild goose chase.[6]

Authentic preferences have a kosher pedigree

Another possibility is that authenticity consists in a certain sort of history. This is a view that – under a different label – has been developed by John Christman in his work on autonomy (see, for examples, Christman 1991, 1993, and 2009: 133–63). Christman (1993: 288) understands autonomy to be a property of an agent in respect of particular preferences at particular times, where whether an agent *A* is autonomous in respect of preference *P* depends on *P*'s genesis satisfying the following four conditions:

- They must not have resisted the development of that desire while they were aware of its development, or would not have resisted it had they been aware;
- That lack of resistance must not have been due to the influence of factors that limit self-reflection;
- That self-reflection must have been minimally rational and must have involved no self-deception;
- The agent must be minimally rational with regard to that desire at the time in question: there must be "no manifest conflicts of desires which significantly affect the agent's behavior."

As I have said, Christman writes about autonomy rather than authenticity. It is clear, however, that – assuming that we are trying to define autonomy because we think it has some connection with our lives going well – his view picks out a conception of authenticity on my terms, because it specifies some (historical) properties which preferences must have.

As we shall see when I set out my own view, I think the conception of authenticity we can find in Christman is not far off the mark. On my view, a preference's authenticity is a matter of whether the true explanation of a preference is one that is hidden from the agent concerned. In many cases, this will depend mostly (or solely) on that preference's history. So, it is possible that my view could imply wholly historical necessary and sufficient conditions for authenticity, of the sort that Christman sets out for autonomy (though I am not entirely sure I would endorse the four conditions Christman suggests, at least not in the form given). I *do* think my conception is preferable to Christman's, but the reason for that is that the explanation-based conception is better motivated. So, I defer discussion of why I finally reject the historical conception of authenticity until my conclusion, once the motivations for my own view have been made clear.

---

[6] Others have expressed concerns about the notion of a "true" self, usually while discussing positive liberty or autonomy. See e.g., Berlin 2002: 181–200, and Berofsky 1995.

The danger of gerrymandering

I could go on, but I think the point is clear. The ease with which we make (some) judgments about which preferences contribute to our lives going well belies the difficulty we find in identifying the properties which make those (and only those) preferences authentic. Setting aside the final (historical) conception, the initial attempts I have surveyed so far all fail because they condemn too much, or allow too much. As the number of failed attempts to capture the distinction stacks up, and we perceive the pattern of their failure (that we have strong convictions about particular preferences which the proposed distinction fails to honor), we might worry that there may after all be no defensible conception of authenticity to be found.

So far, I've been looking for a way of drawing the distinction between authentic and inauthentic preferences that appeal to properties of preferences other than their content. If it turns out, however, that it is in fact always the content of preferences which matters, then authenticity plays no serious role. At best, we might find some gerrymandered property of preferences that is guaranteed to pick out the ones which we already believed were the right ones. But if what makes preferences authentic is their just being the preferences required by some substantive and independent view of what makes someone's life go well, then a theory which incorporates the concept of authenticity will share the unattractive features of the Objective List theories of well-being that I discussed in the first part of my essay. There, I suggested that the concept of authenticity emerges precisely because we want to avoid those features. So, the present danger is that any attempt to find a conception of authenticity will be self-defeating, because it would reveal a hidden commitment to a view on which what makes someone's life go well which is incapable of meeting the first desideratum I set out in my first section.

Of course, that charge will only stick if there is indeed no non-gerrymandered conception of authenticity. I now turn to setting out an account that, I think, does the job.

Authenticity and the Third-person Perspective

To recap: our task is to find a conception of authenticity that can fulfil its purpose by meeting the two desiderata I set out in part one. To do this, it must stand up with adequate robustness to the intuitive scenarios (such as those touched upon in the previous part) which motivate us to look for an account of authenticity in the first place; allow us to distinguish those preferences whose satisfaction contributes to one's life going well; and do so in a way which doesn't turn out to be gerrymandered, which is to say unmotivated save by a covert commitment to a substantive view on which well-being is determined entirely without reference to an agent's preferences.

My proposal is that authentic preferences can be distinguished by a property that they have in virtue of how they are explained. In particular, I suggest the following two definitions:

- **Covert explanation:** An explanation for an agent's preference is *covert* when it is necessarily hidden from that agent.[7]
- **Authenticity**: A preference is inauthentic if its explanation is covert, and authentic if not inauthentic.

In what follows, I elucidate and motivate these two definitions, and explain how they outline a conception of authenticity which satisfies my desiderata without being gerrymandered.

Covert explanation

An explanation for a preference is covert when it is necessarily hidden from the agent concerned. By that, I mean the following:

(a) There is some set of true propositions which, against the background of general laws (for example about psychology or physics), necessitate the preference;[8]
(b) The agent concerned does not believe that set of true propositions;
(c) It could not be the case that (a) is true and (b) is false.

One way that these three conditions could be met is when an agent's preference is explained by a cause which is kept concealed from her. For example, suppose that I am hypnotized in such a way that I want to give all my money to a cult and also remain staunchly ignorant of the true cause. In such a case, (a) (b) and (c) are all true, because if the process is successful (as required by (a)) I will not be aware of it (thus making sure that (b) is true too, and thereby ensuring that (c) holds).

Another way that the three conditions could be met is more complex. In some cases, something's being an explanation of an agent's preference *depends* on its being hidden, either because that preference would cease if the agent came to be aware of the explanation, or because its coming to be known and the preference persisting would just go to show that it hadn't been the right explanation in the first place.

---

[7] This develops an account of covert *influence* I have developed elsewhere, e.g., in Colburn 2011.

[8] This is a deliberately loose version of the model of explanation most famously defended by Carl Hempel (1965: 331–496). It is deliberately loose in that – unlike Hempel – I neither want nor need to specify what sorts of general facts might fit into an explanation. So they might (as he said) just be purely general laws of nature, but they might also be more contextual or pragmatic in character. Crucially, I treat only the particular fact of the explanation, and not the background general laws, as facing the tribunal of experience; it would be implausible to think that a preference is inauthentic if the individual concerned would be inclined to repudiate it only on the basis of a general law rather than anything to do with that preference specifically. My thanks to David Bain, James Humphreys and Chris Snow for pressing this point.

The idea is best illustrated by using a familiar – though apocryphal – story about subliminal advertising in a New Jersey cinema. Imagine the following scenario. During the showing of a film, single-frame adverts for ice cream are flashed on screen.[9] They pass too rapidly to be consciously perceived by members of the audience, but do result in many people in the audience forming a desire to eat ice cream during the interval, which they presumably deem to be based on a proper appreciation of the virtues of ice cream. So the explanation of their preference that an impartial observer would give is hidden from them.

This is not to say that the technique itself is *necessarily* hidden, in the sense that the activity of a secretive hypnotist might be, as Roger Crisp (1987) points out. One can be informed that one has been the subject of subliminal messaging. The point is just that when one is made aware of that, one of the following things must happen:

- One's preference might lapse because one has been told about the technique. ("Actually, come to think of it, I don't really like ice cream after all!") So, there ceases to be a preference to explain once the agent is aware.
- One's preference might persist, but our view of its correct explanation changes: that is, one realizes that one's preference was induced but can adduce other independent reasons for eating ice cream being a good thing to do. If the preference remains genuine then it seems very unlikely that the real reason for it is just that single-frame images of ice cream had been interspersed with her film.[10]
- Finally, one's preference might persist because one is incapable of shaking the induced desire – but this uncomfortable fact leads one to repudiate it. In that case, as Crisp notes, the ostensibly innocent desire for ice cream has become an unwanted craving (1987: 415).

In each case, we can see that the subliminal messaging is covert *insofar as it is the explanation* of a preference: as soon as the agent becomes aware of the purported explanation, either the preference under scrutiny disappears (or turns out to be an unwanted craving), or the preference's persistence shows that the explanation is wrong.

My definition makes reference to an explanation as a set of true claims. This is because, plausibly, most preferences will have complex explanations that involve various elements. In the case of the New Jersey cinema, for example, the subliminal messaging alone cannot explain my new preference for ice cream. A full explanation would also need to refer at least to my having chosen to go to the cinema in the first place, my not being blind, and so on. Not all the elements of an explanation need be covert for the explanation to be covert as a

---

[9] This familiar case is based on a report (Packard 1957) of an experiment supposedly carried out by a marketing researcher called James Vickery. Vickery subsequently admitted that he had faked the evidence (Danzig 1962), but the (now admittedly imaginary) case remains a useful one for highlighting problematic features of preference-formation mechanisms.

[10] Of course, they might. But in that case I suppose it would just turn out that for some unusual people subliminal messaging need not make the induced preferences inauthentic.

whole. There is (or at least need be) nothing hidden or non-transparent to me about my having chosen to visit the cinema, and it would be surprising if I were incapable of noticing that I am not blind. The point is that condition (b) above would be false only if an agent believes *all* of the true propositions that form the explanation of her belief. So, condition (c) is met so long as there is *some* true proposition which is part of the explanation which the agent couldn't believe alongside the preference's persisting.

So, to recap, on my definition, authenticity consists in a preference's not having a covert explanation. In what follows, I offer three reasons to think this a good conception of authenticity.

Fidelity to the paradigm cases

My conception makes good sense of paradigm cases of inauthenticity, such as those discussed in the previous section and at the start of part two. The problem with hypnosis and subliminal messaging, to repeat, is that they are mechanisms that ensure that preference formation takes place (in Elster's words) "behind the back" of the agent concerned, and hence have explanations which are not transparent to the agent (1983: 117). My notion of covert explanation gives robust content to this plausible but vague observation. It also reveals the feature such cases share with other adaptive preference formation. That phenomenon involves someone's sincere preferences altering in response to the unavailability of certain options, but where that does *not* seem to them the explanation for their preferences. (In Aesop and La Bruyere's fable, which Elster uses to illustrate his discussion, the fox that finds that she cannot after all reach the bunch of grapes she had desired says to herself "Those grapes were sour, anyway!") Though Elster does not say so, it is precisely this feature which distinguishes adaptation from related but generally innocuous phenomena whereby people prioritize amongst their ambitions in light of their likely success in achieving them, or consciously seek to shape their own preferences so as to better fit with their own talents and circumstances.[11]

Between Scylla and Charybdis

At the beginning of this part of my discussion, I set out the following two desiderata: first, that someone's own perspective or preferences should be central in determining what counts as their life going well, and secondly that not all manifest preference-satisfaction contributes to their well-being. There, I suggested that the plausibility of these two theses explains why we have a concept of authenticity in the first place; so, any conception of authenticity must meet them if it is not to be a failure. Mine does. It allows that one's authentic preferences – those which don't have covert explanations – can determine what makes one's life go well.

---

[11] See Colburn 2011 for further discussion of this point. I say that the other coping strategies are "generally" innocuous because they might still be problematic in cases where people's likely success or circumstances are externally limited in some way that we think unjust. The point is that such concerns are not directly to do with authenticity.

But it does rule out a class of manifest preferences – namely those which have covert explanations – thereby giving us the critical resources necessary to provide distance from the Manifest Preference Satisfaction view.

Thus far, I haven't said much about how many manifest preferences will be judged inauthentic by this conception. Many of the mechanisms I have considered so far – hypnosis and subliminal messaging, for example – have been deliberately far-fetched, and might contribute to the impression that in practice my conception of authenticity is relatively undemanding: we are unlikely to encounter such phenomena outwith the realms of science fiction. So in practice my conception might have little bite; and, more worryingly, we might think that in fact there is very little to separate mine from the Manifest Preference Satisfaction view, which it seeks to avoid.

There are two reasons to think that my conception will have more substantive practical force than the focus on spectacular examples might suggest. For one thing, we might worry (as Crisp does) that the troubling characteristics of subliminal messaging are also present in a lot of ostensibly less problematic persuasive advertising. (Crisp 1987: 415–6) To the extent that such advertising plays an important role in shaping many people's preferences, Crisp's point would imply that the saturation of the modern world by such advertising leads to widespread inauthenticity. For another, many have argued that pervasive adaptive preference formation is one of the key mechanisms that perpetuate inequalities based on gender,[12] race[13] and class.[14] If those arguments are sound, inauthenticity will turn out to be an endemic problem intimately bound up with the central concerns of liberal political philosophy.

Authenticity and the third-person perspective

In the second part of my discussion, I noted that some ways of characterising authenticity gerrymander the concept, by defining it in a way guaranteed to identify a particular set of preferences prejudged to be valuable because of their content. Such theories of authenticity are self-defeating, as well as smacking of intellectual dishonesty. Does my conception avoid the charge of gerrymandering?

---

[12] See, for example, Babbitt (1995) and Nussbaum (2001).

[13] For example, Kwame Anthony Appiah (1994) has argued that the essentialist conceptions of race embodied in some black liberation movements can be as malign for some individuals as the racist oppression they seek to oppose. Many of his examples hinge on the way that such conceptions of race can covertly alter people's preferences and expectations for themselves – that is to say, adaptively (though Appiah does not exactly put it that way).

[14] For example, Amartya Sen (1999), though his focus is more on poverty than class; and writers on working-class participation in higher education (e.g., Bridges 2006 and Watts 2009).

At first sight, it looks innocent. So long as a preference does not have a covert explanation, it counts as authentic by my definition. That potentially makes my conception extremely catholic. It does not, for example, rule out people having stupid, irrational or malicious preferences which are nevertheless authentic, and whose satisfaction might therefore contribute to their well-being.[15] It is possible that there are some things which, as a matter of fact, nobody could prefer authentically. I tend to think not: like Wilhelm von Humboldt, I suspect that "there is no pursuit whatever that many not be ennobling and give to human nature some worthy and determinate form" (1969: 28–29). Whether that is right or wrong, however, the point is that any necessarily inauthentic preferences will only be identified *post hoc*, without their exclusion being built into the conception from the start because we're secretly committed to an Objective List theory of well-being.

Nevertheless, we might still worry that there's something problematic about my conception of authenticity, because (like any such conception) it faces the following dilemma. Unless it is independently motivated, it looks like an *ad hoc* attempt to subsume the various phenomena described at the start of part two under one conception; but if it *is* independently motivated, that independent motivation looks like it is doing the real normative work, and I am once more vulnerable to the charge of gerrymandering. To show that this argument fails, I now turn to showing that my conception of authenticity is neither ill motivated nor self-defeating. In so doing, I set out why I think my conception is uniquely well-placed to meet the desiderata established by reflecting on why we have a concept of authenticity in the first place.

Let us reflect on what is different about preferences which have covert explanations. Suppose that we are trying to account (for instance) for my desire to learn to play the piano. If I am asked to explain that preference, then I will generally have some story to hand: perhaps I might say "Because I chose to devote myself to learning to play the piano years ago, and it's important to me that I fulfil that ambition"; or "Because piano music is valuable." In normal circumstances, an ideally placed third-person observer would echo my answers if she were trying to explain my preferences. She would say "Because he wants to fulfil his chosen ambition," or "Because he believes that piano music is valuable." And that is as it should be. When thinking about *why* someone has the preferences that they do, their own perspective on what is important to them has some sort of authority.

The problem with cases of covert explanation is that the first- and third-person explanations of preferences come apart. So the fox tells us that he doesn't desire to eat the grapes because they're sour; Aesop explains it by reminding us that the fox couldn't jump high enough. The hypnotized cult member says that she is handing over her money because she feels like it (and because Zeus is great); we know that it is down to the hypnotic techniques deployed by the cult leader which our hapless member has forgotten. The working class student explains that he doesn't want to go to university because there's no point to learning for its own sake;

---

[15] This isn't to say that we should allow people to satisfy malicious preferences, nor that we mightn't have reasons to ensure through education that people's preferences are not determined by ignorance, stupidity or irrationality.

his teachers can tell us that his lack of ambition comes from incomprehension and lack of ambition on the part of his parents. In general, then, when the explanation for a preference is covert, the reasons that an individual adduces for her commitment, however sincerely, are not the "real" reasons that a third-party perspective would reveal. She has become opaque to herself.

The crucial thing is that a concern for avoiding such opacity looks like it's motivated by one and the same thing as the theory of authenticity in the first place, namely a recognition an individual's perspective (as reflected in her preferences) has a certain authority over what makes her life go well. That authority is subverted when that individual's perspective is deposed, as it is when her preferences are covertly explained. So, *if* we are in the first place inclined to accept the first desideratum as stated in part one, then we can also generate the principle which provides the critical distance demanded by the second desideratum. That is to say, it is precisely our concern for the authority of the first-person perspective when it comes to questions about what makes someone's life go well that forces us to think that something goes awry when the explanation someone gives from that perspective deviates from the true explanation as given by an ideally-placed third person.

It is a question for a different time precisely what values might underwrite this principled focus on people's first-personal perceptions on what makes their lives go well. Elsewhere, I have defended a conception of autonomy on which that value is understood to consist in an individual deciding for herself what is valuable, and living her life in accordance with that decision (Colburn 2010a). A prior commitment to a value of autonomy of this sort (whether or not that is conceived as a component of our theory of well-being) would be sufficient. But there may be other foundations too, so long as (like autonomy on my view) they take there to be something centrally important about an individual's own perspective on her life.[16] My point here is just this: anything which is sufficient to motivate the two desiderata outlined in part one, and hence make us search for a conception of authenticity in the first place, will be at one and the same time sufficient for motivating the particular conception of authenticity that I defend here, without smuggling in any additional (and self-defeating) normative foundations.

---

[16] It might, of course, be that autonomy is the only value that will do the job. Elsewhere, I have argued that the value of autonomy has an unusual internal structure, in virtue of which it allows an individual's own judgments to be crucial in determining what states of affairs are valuable or not (see e.g., Colburn 2010b). If autonomy *is* the only value we might appeal to here, then accepting my conception of authenticity would also commit one to recognising a substantial value of individual autonomy. Indeed, if I am right in my discussion of the function that the concept of authenticity is supposed to play, then accepting even that more general claim will have the same effect. Someone who completely rejects the notion that autonomy is valuable will, therefore, not want to accept my theory of authenticity; but in light of what I have said in this essay, it seems as though such a person would need then to abandon the concept of authenticity entirely. My thanks to Iona Stevenson for pushing me to think about this connection.

Conclusion

In this chapter, I have defended definitions both of the *concept* of authenticity and of a particular *conception* of authenticity. Concerning the first, I argued that the concept arises when we seek to understand the relationship between someone's preferences and their life going well; in particular, the concept points to a local property of (some) preferences in virtue of which their satisfaction contributes to that global goal. Concerning the second, I proposed a conception of authenticity on which that property consists in a preference's not having a covert explanation. This, I argued, is a better account than its rivals for three reasons: it fits with intuitive paradigm cases; it meets the desiderata which spurred our search for a conception of authenticity in the first place; and it does so in a way which doesn't make the theory self-defeating. In this respect, I suggested, it is uniquely well placed. If that's right, then if there's any call for a conception of authenticity at all, this is the right one to adopt.

Before finishing, it is worth making plain the difference between my conception and the one rival conception I left open in part two of my discussion. There, I noted that my conception might be coextensive with something like Christman's historical conception of autonomy, and that the arguments I have deployed against my other rivals – namely that they either rely on implausible assumptions or draw the line between authenticity and inauthenticity in implausible places – probably don't work here.

The problem with the historical conception is that it fails to identify why the history of a preference matters. Why, we might ask, is it important that we are rational in respect of a preference, or would not have resisted its formation had we been aware of it? Merely setting out a detailed set of historical conditions which we take to be severally necessary and jointly sufficient for authenticity does not answer those fundamental questions. But if those questions are not answered, the conception of authenticity that results will have very shallow foundations indeed, even if it happens to give the right answers in all the cases we care about. At the very least, then, there is a lacuna in such accounts where there ought to be a statement of the ideal which explains why *that* set of necessary and sufficient conditions captures something important.

My concluding reflections in part three on the first-person perspective offer one possible way of filling that gap. On my view, a preference's history is important, not in itself, but because it can (maybe must) determine the relationship between an agent and her preferences. Ultimately, however, the thing that matters is that relationship's being a good and transparent one, rather than involving opacity and covertness. Still, this does offer a reason to think a preference's history very important indeed, and would suffice to motivate some historical necessary and sufficient conditions on authenticity.

For what it's worth, I think that this proposal is compatible with the revised version of his view that Christman develops in this volume (Chapter 2). Christman's reflections there on the complex ways in which an agent's self-conception interacts with her preferences, and his emphasis on the importance of the narratives we deploy when constructing our practical identities, all seem to me extremely important for a fully fleshed-out understanding both of autonomy and of authenticity. Their being so, however, gives us good reason to think that

the bare historical facts don't count for nearly as much as Christman's earlier view suggested. My theory – which focuses on the broader category of *explanation*, rather than *history*, and which focuses on opacity of explanation as the criterion for inauthenticity – is well-placed to accommodate Christman's reflections by showing how the historical facts play an important but indirect role, insofar as they will be part (but not the whole) of the explanatory story which constitutes an individual's understanding of herself and her ambitions.

This eirenic conclusion may satisfy the defender of the historical conception; but it comes, of course, at the cost of abandoning the attempt to defend that conception as a clear and distinct alternative way of construing authenticity (or at any rate as a reason for rejecting the theory I present here). If that cost is too high, I can see two alternative strategies for the defender of the historical conception to deploy. One would be to identify some *other* substantive theory of well-being that might explain why the purportedly valuable history is a good one for preferences to have. That strategy bears a burden of proof, and runs the risk of the gerrymandering I condemned in part two.

The other strategy would be simply to insist that it *just is* valuable for our preferences to have a history of this sort rather than another. That sounds unlikely – is this *really* the sort of place where we can fairly claim to have hit moral bedrock? – and is open to the charge that it fetishizes something which is really only important as a means to an end.

I cannot here rule out the possibility that the defender of the historical conception might deploy one of these strategies and try thereby to respond to these charges: the preceding comments are, after all, merely indications of the theoretical costs of that position rather than outright refutations. The costs are high enough, however, that I predict that few will seek to pay it. And if that is so, my conception of authenticity as absence of covert explanation is vindicated.

References

Appiah, Kwame Anthony. 1994. "Race, culture, identity"; available at: http://www.tannerlectures.utah.edu/lectures/documents/Appiah96.pdf

Arneson, Richard J. 1999. "Human Flourishing Versus Desire Satisfaction," *Social Philosophy & Policy* 16: 113–142.

Babbitt, Susan. 1995. *Impossible Dreams: Rationality, Integrity, and Moral Imagination.* Boulder: Westview.

Berlin, Isaiah. 2002. "Two Concepts of Liberty." In *Liberty,* edited by H. Hardy. Oxford: Oxford University Press, pp. 166–217.

Berofsky, Bernard. 1995. *Liberation from Self*. New York: Cambridge University Press.

Bridges, David. 2006. "Adaptive Preference, Justice and Identity in the Context of Widening Participation in Higher Education," *Ethics and Education* 1: 15–28.

Christman, John. 1991. "Autonomy and Personal History," *Canadian Journal of Philosophy* 21: 1–24.

—. 1993. "Defending Historical Autonomy: A Reply to Professor Mele," *Canadian Journal of Philosophy* 23: 281–289.

—. 2009. *The Politics of Persons*. Cambridge: Cambridge University Press.

Colburn, Ben. 2010a. *Autonomy and Liberalism*. New York: Routledge.

—. 2010b. "Anti-perfectionisms and Autonomy," *Analysis* 70: 247–255.

—. 2011. "Autonomy and Adaptive Preferences," *Utilitas* 23: 52–71.

Craig. Edward. 1990. *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Oxford University Press.

Crisp, Roger. 1987. "Persuasive Advertising, Autonomy, and the Creation of Desire," *Journal of Business Ethics* 6: 413–418.

—. 2006. *Reasons and the Good*. Oxford: Oxford University Press.

Danzig, Fred. 1962. "Subliminal advertising – Today it's just historic flashback for researcher Vicary," *Advertising Age,* 17 September.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Dworkin, Ronald. 2012. *Justice for Hedgehogs*. Cambridge MA: Harvard University Press.

Elster, Jon. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Feldman, Fred. 2004. *Pleasure and the Good Life*. Oxford: Oxford University Press.

Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

von Humboldt, Wilhelm. 1969. *The Limits of State Action*, ed. and trans. J.W. Burrow. Cambridge: Cambridge University Press. Originally published 1810 as *Ideen zu einem Versuch, die Gränzen der Wirksamkeit des Staats zu bestimmen*.

Kierkegaard, Søren. 1983. *Fear and Trembling*. Princeton: Princeton University Press, translated H.V. & E. H. Hong. Originally published 1843.

Nussbaum, Martha. 2001. "Adaptive Preferences and Women's Options," *Economics and Philosophy* 17: 67–88.

Olsaretti, Serena, and Richard J. Arneson (eds). 2006. *Preferences and Well-Being*. Cambridge: Cambridge University Press.

Packard, Vance. 1957. *The Hidden Persuaders*. New York: Pocket Books.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Sandven, Tore. 1995. "Intentional Action and Pure Causality: A Critical Discussion of Some Central Conceptual Distinctions in the Work of Jon Elster," *Philosophy of the Social Sciences* 25: 286–317.

Sartre, Jean-Paul. 1948. *Being and Nothingness*. New York: Philosophical Library, translated H.E. Barnes. Originally published 1943.

—. 1992. *Notebooks for an Ethics*, translated by D. Pellauer. Chicago: University of Chicago Press. Originally published 1983.

Sen, Amartya. 1999. *Development as Freedom*. Oxford: Oxford University Press.

Sumner, L.W. 1999. *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press.

Trilling, Lionel. 1972. *Sincerity and Authenticity*. Cambridge MA: Harvard University Press.

Watts, Michael. 2009. "Sen and the Art of Motorcycle Maintenance: Adaptive Preferences and Higher Education," *Studies in Philosophy and Education* 28: 425–436.

Williams, Bernard. 2004. *Truth and Truthfulness: An Essay in Genealogy*. Princeton: Princeton University Press.