n

Colburn, B. (2011) Autonomy and adaptive preferences. Utilitas, 23(1), pp. 52-71.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/44230/

Deposited on: 15 February 2016

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Autonomy and End of Life Decisions: A Paradox

Ben Colburn, University of Glasgow

Published in *Utilitas* 23 (2011): 52-71.

ABSTRACT

Adaptive preference formation is the unconscious altering of our preferences in light of the options we have available. Jon Elster has argued that this is bad because it undermines our autonomy. I agree, but think that Elster's explanation of *why* is lacking. So, I draw on a richer account of autonomy to give the following answer. Preferences formed through adaptation are characterised by *covert influence* (that is, explanations of which an agent herself is necessarily unaware), and covert influence undermines our autonomy because it undermines the extent to which an agent's preferences are ones that she has decided upon for herself. This answer fills the lacuna in Elster's argument. It also allows us to draw a principled distinction between adaptive preference formation and the closely related – but potentially autonomy-enhancing – phenomenon of character planning.

INTRODUCTION

Adaptive preference formation – that is, the unconscious altering of our preferences in light of the options we have available – is often thought problematic, on various grounds. One suggestion – made most influentially by Jon Elster – is that adaptive preference formation is bad because it affects our *autonomy*. Elster's argument to that end, however, is hampered by two problems. First, it is not clear what notion of autonomy he has in mind, nor indeed whether it is really *autonomy* (as opposed to rationality) that he cares about. For that reason, he is unable to account for the badness of adaptive preference formation Secondly, Elster is unable to offer a principled distinction between that phenomenon (which is supposed to be bad) and conscious character formation (which isn't). In this paper, I offer a better account than Elster's. By drawing on a richer account of autonomy, I show that adaptive preference formation is bad because it compromises the independence of our commitments: preferences formed through adaptation are characterised by *covert influence* (that is, explanations of which an agent herself is necessarily unaware). This also allows us to draw the necessary distinction with conscious character planning. While adaptive preference formation is always covert, character planning never is, and this explains why the latter can be positively supportive of our autonomy.

1. ADAPTIVE PREFERENCE FORMATION

Jon Elster has famously analysed various different mechanisms whereby the rationality of our preferences can be subverted.[1] The most frequently discussed such mechanism is one that affects the formation or change of preferences, namely *adaptive preference formation*. Preferences that are formed in this way involve an element of adaptation to circumstances: our having we have the

[1] J. Elster, *Sour Grapes: Studies in the Subversion of Rationality* (Cambridge, 1983).

preferences we do is explained by our beliefs about the unavailability of certain options, rather than (for example) the intrinsic qualities of the options we *do* have, and for which we have formed a preference. Elster illustrates the phenomenon by evoking Aesop and La Fontaine's fable of the Fox and the Grapes. In that story, a fox sees some grapes hanging on a vine, but cannot reach them. So, the fox says 'Those grapes are sour, anyway!', and loses the preference for the grapes that he had before he realised that eating them was not a real option for him. His 'sour grapes' reasoning is the mechanism whereby his preferences are adapted in response to the constraints placed on his option set.[2]

Elster's contention is that a preference is problematic when formed by such a mechanism. There are various reasons for this. One, for example, is that adaptive preferences subvert an agent's rationality. More interestingly, though, Elster says that such cases pose a problem for an agent's autonomy.[3] If that's right, then it shows how adaptive preference formation is a phenomenon with interest beyond a mere analysis of rationality. The notion of autonomy – disputed and unclear though it is – plays a role in a wide variety of moral and political arguments. If it turns out that adaptive preference formation is a mechanism which systematically undermines autonomy, then moral and political philosophers need to know why, and how to avoid it.

2. ELSTER'S CONCEPTION OF AUTONOMY

This potentially important normative payoff makes it frustrating that Elster's arguments about the connection between adaptive preferences and autonomy are so unclear. This is for two reasons: first, it is not clear that Elster has in mind an ideal of *autonomy* (rather than just a conception of *rationality*), and secondly, even if we think that there is a distinct appeal to autonomy, it is unclear what autonomy amounts to on his view.

Elster describes autonomy as 'substantive rationality of desires ... being for desires what judgment is for belief'.[4] Now, if Elster had given us a definition of substantive rationality, then the game would be up: by 'autonomy' he would just mean 'whatever is required of desires (as opposed to beliefs) for them to fall under this broader category of substantively rational mental states', and we would then be looking rather at an ideal of rationality than of autonomy as it is discussed by moral and political philosophers.

As it happens, though, I think this is not the right interpretation, as is clear when we ask what is meant by 'substantive rationality'. Elster contrasts it to 'thin rationality' – which requires only consistency in our mental states – but gives no general definition, beyond noting that our everyday use of the term 'rational' requires something that goes 'beyond the excusively formal considerations' of consistency.[5] Rather than give such a definition, Elster goes through different types of mental state and explains what rationality in this 'more substantive sense' requires. For example, to be substantively rational, beliefs must be 'grounded in the available evidence', which

---

[2] Elster, *Sour Grapes*, p. 109.

[3] Elster, *Sour Grapes*, p. 20. Others have made the same claim, e.g. John Christman in 'Autonomy and Personal History', *Canadian Journal of Philosophy* 21 (1991): 1-24; and David Zimmerman in 'Making do: Troubling Stoic Tendencies in an Otherwise Compelling Theory of Autonomy', *Canadian Journal of Philosophy* 30 (2000): 25-54, esp. 27-30.

[4] Elster, *Sour Grapes*, p. 30.

[5] Elster, *Sour Grapes*, p. 15.

is to say the output of a process of good epistemic *judgement*.[6] Autonomy is the analogous criterion for desires. So, it rather looks as though Elster uses the term 'substantive rationality' as an umbrella term, designed to capture the various normative – but perhaps non-moral – standards by which we judge mental states. The standard that beliefs must live up to is that of formation through sound judgment on the basis of the available evidence. The standard that desires must live up to is that of autonomy. And so on.

Let us assume that this is the right reading. This shows that the question of what autonomy consists in is still a live one. Elster may still want to link autonomy to rationality in some sense, but the crucial point is that a definition of autonomy should be prior to one of substantive rationality. Moreover, we might accept a proposed normative standard for desire-formation without our interest in such a standard coming from a belief that it tells us anything interesting about rationality.

Frustratingly, at precisely the point where we might want a definition, Elster admits defeat. He runs through various possibilities – implicit definition by pointing out 'persons that apparently are in control over the processes whereby their desires are formed', or explicit definitions like 'autonomy desires ... have been deliberately chosen, acquired or modified' – but rejects all as unsatisfactory for different reasons.[7] So, he falls back on the more modest aim of running through some crucial cases in which autonomy is undermined through our desires being formed by questionable mechanisms, and hoping that that will help us discover what autonomy is. Indeed, he suggests that this is enough for his purposes, saying that

> In the present work, autonomy will have to be understood as a mere residual, as what is left after we have eliminated the desires that have been shaped by one of the mechanisms on the short list for irrational preference-formation.[8]

Unfortunately for Elster, that can't be satisfactory. We need a positive account of autonomy – not a 'mere residual' – if we are to know which mechanisms are unsatisfactory and why. Admittedly, the passage just quoted doesn't imply that no such account is possible, nor indeed that we mightn't find it by reflecting on the quality of desires left after various uncontroversially unsatisfactory mechanisms have been eliminated. But that means that everything hinges on the question of *why* adaptive preferences are bad, and our being able to have some grip on the answer before we know what autonomy is. So, we seem to be trapped in a circle. Elster can neither explain the badness of adaptive preferences nor help us discover what autonomy is by using the implicit definition by residue that he espouses.

3. THE CONTRAST WITH CHARACTER FORMATION

In his discussion of adaptive preferences, Elster distinguishes the sour grapes mechanism with which he is particularly concerned from various related phenomena: counteradaptive preferences, precommitment, addiction and so on.[9] For the most part these are diagnosed as

[6] Elster, *Sour Grapes*, pp. 15-17.
[7] Elster, *Sour Grapes*, pp. 21-22.
[8] Elster, *Sour Grapes*, p. 24.
[9] Elster, *Sour Grapes*, pp. 111-24.

being problematic, but not necessarily for the same reason as adaptive preferences. One phenomenon he mentions is *not* intended to be problematic, and this is conscious character planning – that is, being aware of the limitations in one's options and moulding one's projects and inclinations so as to settle on preferences which one can fulfil.[10]. Such planning, Elster says, is a good thing from the point of view of autonomy; or at any rate, if it *is* bad, it's not bad for the same reasons as adaptive preference formation.

The problem is that, in some respects, character planning and adaptive preference formation look extremely similar. Both involve an agent's preferences changing (or being formed by) their beliefs about the limitations of their option sets. So, if we want to give different moral appraisals of the two phenomena, we must be able to point to a sharp and principled distinction between them. Elster fails to do this, because he characterises the difference between character planning and adaptive preference formation in several non co-extensive ways.

Some of these are plainly intended to be descriptive, rather than definitive. So, for example, Elster says that adaptive preference formation tends to 'overshoot' what is determined by one's possibilities (meaning that preferences are modified more than is strictly required), whereas character planning can 'shape one's wants so as to coincide exactly with ... one's possibilities'; and notes that the former usually involves downgrading inaccessible options and the latter involves upgrading accessible ones.[11] It would seem uncharitable, though, to read Elster as saying that these contrasts are what the distinction itself consists in, although others who have worried about adaptive preference formation do seem guilty of this error.[12] Elster does say enough, though, for us to identify three different proposals for drawing the crucial distinction. As we will see in §6, I think that on each reading he identifies something crucial, but – lacking as he does the unifying conception of autonomy I will introduce in §4 – each reading proves unsatisfactory.[13]

A: Causal vs consciously engineered

In one place, Elster characterises adaptive preference formation as a 'purely causal process' and contrasts it with character planning as 'engineered by conscious strategies of liberation'.[14] In another paper, Elster says that the problematic feature of adaptive preference formation is that 'the source of the preference change is not in the person', whereas, by implication, the source of

[10] Elster, *Sour Grapes*, pp. 117-19.

[11] Elster, *Sour Grapes*, pp. 118-19.

[12] e.g. M. Rickard, 'Sour-grapes, Rational Desires and Objective Consequentialism', *Philosophical Studies* 80 (199): 279-303, at 284.

[13] Others besides Elster have tried to characterise the distinction. For the most part their distinctions tend to map onto one or other of the proposals for interpreting Elster that I discuss here, so I do not mention them separately. One exception is Luc Bovens, who says that the two types of phenomenon differ differ in respect of the semantic content of the preferences we end up with: adaptive preference formation involves adjusting one's preference for *tokens* without engaging in reasoning about the desirability of *types*, whereas 'a typical case of character planning is the more involved project in which I can adjust my *reasons* for the ranking at hand'. See L. Bovens 'Sour Grapes and Character Planning', *The Journal of Philosophy* 89 (1992): 57-78, at 74. I do not consider Bovens's proposal here, for the same reasons as those given by Zimmerman, who complains that its focus on the content of preferences is misplaced, and leads Bovens to ignore some important variants of adaptive preference formation (see his 'Sour grapes, self-abnegation and character building', *The Monist* 86 (2003): 220-41, at 228-35.

[14] Elster, *Sour Grapes*, p. 117.

preference change in the case of character planning *is*.[15] Such statements might imply that the distinction maps roughly onto a causal/non-causal divide. However, this is implausible. The presence or absence of causation can hardly be what is at issue. Unless Elster wants to defend the view that character planning allows us to slip the shackles of physical determinism (a controversial metaphysical thesis for which he offers no argument), any sense in which adaptive preference formation is 'purely causal' must also be one in which character planning is too.[16]

B: Unconscious vs conscious

Perhaps the relevant feature of character planning is that it is *conscious*. If that is so, Elster's reference to a 'purely causal process' might be read as an oblique claim that adaptive preference formation is typically *unconscious* – in his words it takes place 'behind the back of the agent concerned'.[17] This seems more plausible than the causal/non-causal contrast. However, it can't be what Elster is after either. Recall that we need a distinction which can ground Elster's claim that adaptive preference formation is bad and character formation is not. But there are many processes of preference formation that are unconscious – indeed, we might think that *most* preferences are formed unconsciously, with conscious character planning being something of a rarity. The desire for food is not normally induced through conscious hunger-creation. A preference for sleep is only rarely something which someone has consciously to cultivate at the end of the day. It would be a very austere notion of autonomy indeed which judged that eating and sleeping were, under almost all circumstances, problematic from the point of view of autonomy. That would follow, though, from thinking that it is the mere fact of adaptive preference formation being *unconscious* that distinguished it from morally unproblematic character planning.

C: Drives versus meta-preferences

Elster claims that the distinction between the two phenomena is 'the difference between preferences being shaped by drives or by meta-preferences.'[18] Elster must mean one of two things by 'drives': either he means 'first-order' (as opposed to higher-order, or 'meta-') preferences, or he means some rank of preferences which is lower than what we usually refer to as 'first-order'. It doesn't really matter which. The crucial point is that on *this* proposal, the difference between character planning and adaptive preference formation is that the latter involves lower-order preferences being shaped by higher-order ones, and the former does not.

This is a different distinction to both *A* and *B*. 'Drives' and 'meta-preferences' are presumably both mental states, and so whatever our view on the role of causation in the mental we will end up classifying them on the same side of the causal/non-causal divide. Moreover,

---

[15] Elster, *Sour Grapes*, pp. 109-10.

[16] For further discussion of Elster's distinction construed this way, see Tore Sandven in '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 (1995): 286-317; 'Autonomy, adaptation, and rationality – a critical discussion of Jon Elster's concept of "sour grapes" Part I', *Philosophy of the Social Sciences* 29 (1999): 3-31; and 'Autonomy, adaptation, and rationality – a critical discussion of Jon Elster's concept of "sour grapes" Part II', *Philosophy of the Social Sciences* 29 (1999): 173-205.

[17] Elster, *Sour Grapes*, p. 117. See also Zimmerman 'Sour grapes', 221.

[18] Elster, *Sour Grapes*, p. 117.

there is no reason to think that the process by which our preferences are shaped by meta-preferences is necessarily conscious, nor either the converse. For a counterexample to the former, consider Marilyn Friedman's case of a oppressed spouse whose higher-order preference to have fully obedient desires leads to her unconsciously suppressing her first-order preference not to wash the dishes.[19] For a counterexample to the latter, imagine someone who is perpetually and powerfully hungry, and so consciously cultivates preferences for cheap victuals so that she might get as much food as possible.

In what follows, I analyse proposal *C* in much more detail than either *A* or *B*. This is because at first sight it looks much like a much more promising way of construing the distinction between adaptive preference formation and character planning. For one thing, it is clear, assuming that one thinks that the hierarchical model of preferences – originally proposed by Harry Frankfurt – is correct.[20] For another, it suggests that Elster might be able to appeal, in responding to my worries in §2, to the influential conception of autonomy which stems from Frankfurt's model, on which autonomy consists in higher-order endorsement of lower-order preferences. Gerald Dworkin defines it thus:

> Autonomy is conceived of as a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values.[21]

Now, if he wanted to co-opt Dworkin model of autonomy to explain the difference between adaptive preference formation and character planning, Elster would needs to modify things somewhat. For one thing, he would have to say that the former is not bad *just* because it involves lower-order preferences influencing each other – there doesn't seem anything wrong with that, and such processes are on Dworkin's account neutral vis-à-vis autonomy. Rather, he would have to say that adaptive preference formation is bad because it involves one's first-order preferences having the shape they do despite the fact that, if we reflected on them in light of our second-order preferences, we would repudiate them.

I think that using a theory of autonomy to explain the distinction between the two phenomena is the right approach, as I show in §6. However, Dworkin's conception of autonomy will not do the work that Elster needs it to do, and for that reason the proposal based on higher-order endorsement fails.

The problem for Elster here is this. The motivation and content for the distinction between adaptive preference formation and character planning now comes from Dworkin's conception of autonomy. So, Elster's account stands or falls with Dworkin's, and is vulnerable to the significant criticism that the latter has attracted. For example, Gary Watson and Irving Thalberg influentially complain that the crucial notion of identification is sufficiently vague to

---

[19] M. Friedman 'Autonomy and the split-level self', *Southern Journal of Philosophy* 24 (1986): 19-35.

[20] H. Frankfurt 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* 68 (1971): 5-20.

[21] G. Dworkin *The Theory and Practice of Autonomy* (Cambridge 1988): p. 20. Dworkin's book was published after Elster's, so the latter can't have had in mind the precise formulation just quoted. However, Dworkin expressed a broadly similar idea earlier, e.g. in 'Autonomy and Behaviour Control', *Hastings Centre Report* 6 (1976): 23-28; and 'The Concept of Autonomy', in R. Haller ed. *Science and Ethics.* (Amsterdam, 1981): pp. 203-13.

make one sceptical about the whole theory.[22] I shall not here discuss most criticisms in detail, since their main relevance is to show that Elster, if he is to rely on Dworkin's theory of autonomy, carries a significant burden of proof which is as yet undischarged. That is enough to motivate someone who sympathises with Elster to prefer the account I sketch in the next few sections. However, one line of attack *is* relevant, since it both threatens Elster's moral assessment of the distinction between adaptive preference formation and character planning, and motivates my preferring the conception of autonomy I set out in §§4-5.

Above, I mentioned a powerful point made by Marilyn Friedman. Friedman suggests that there are cases of conflict between higher- and lower-order preferences where, contrary to Dworkin's theory, an individual is more autonomous if she acts on the latter and attempts to revise the former.[23] For example, she asks us to consider a woman who has been brought up to desire some oppressive level of obedience to her husband. Such an individual might have a strong first-order preference not to wash the dishes, and a strong higher-order preference not to have such disobedient preferences. From the point of view of autonomy, Friedman points out that it is not at all clear that the first-order preferences should be overridden. Indeed, our intuitions rather favour the opposite, and insofar as Dworkin must advocate the first course, his theory is implausible.

In response to criticisms like Friedman's, the debate over Dworkin's conception has become somewhat stuck in a baroque fugue between critics (who propose cases as counterexamples to the conception) and proponents (who offer small modifications to address each counterexample as it arrives). The debate is inconclusive, and its details need not concern us here.[24] Two points only need to be made.

The first is that Friedman's example hinges on worries about the provenance of people's higher-order attitudes. Our obedient housewife's higher-order desire to be obedient is questionable because we think she has been brainwashed into it. Merely being higher-order doesn't guarantee that it is unproblematic from the point of view of autonomy. Indeed, we can easily give a more detailed description of Friedman's case so that the higher-order preference is itself the result of what looks like adaptive preference formation. Perhaps the reason that the housewife has such a strong preference to have only obedient preferences is as a reaction to the limited options available for an independent-minded woman in a chauvinistic society. That seems eminently possible – or at any rate, whether it *is* is a matter for psychological investigation, not analytic reflection. However, it is ruled out analytically by the Dworkin-inspired way of capturing the distinction between adaptive preference-formation and character planning: the process where lower-order preferences are brought into line with higher-order ones is by definition the latter and not the former. I assume that this is sufficiently implausible as to rule out the initially promising proposal that we understand Elster's distinction as piggybacking on Dworkin's hierarchical account of autonomy.

The second point is broader, and returns to the issue I first raised in §2. We turned to Dworkin to give us a suitable conception of autonomy that could be used to underwrite both

[22] G. Watson, 'Free Agency', *Journal of Philosophy* 72 (1975): 205-20; I. Thalberg 'Hierarchical Analyses of Unfree Action', *Canadian Journal of Philosophy* 8 (1978): 211-26.

[23] Friedman, 'Autonomy and the split-level self'.

[24] There are more such arguments in Friedman 'Autonomy and the split-level self'; Thalberg 'Hierarchical Analyses', and M. Oshana 'How much should we value autonomy?', *Social Philosophy and Policy* 20 (2) (2003): 99-126. Some defences can be found in M. Bratman 'Autonomy and Hierarchy', *Social Philosophy and Policy* 20 (2) (2003): 156-76.

Elster's claims about the badness of adaptive preference formation and also his distinction between that phenomenon and character formation. The fact that Dworkin's conception has become enmired in such a fruitless debate reveals, I think, a deeper worry about that way of understanding autonomy. Even if we can provide endless *ad hoc* modifications in response to cases like Friedman's, we can still ask: Why should we think it valuable for people to have their hierarchy of attitudes arranged in the particular way he describes? Why should higher-order attitudes be authoritative? Insofar as the hierarchical theory leaves such questions open, it is inconclusive, and seems most likely itself tacitly to rely on a different conception of autonomy which is actually doing the normative work. Since uncovering such a conception will help rescue Elster as well, I now turn to setting out my positive account.

4. AUTONOMY REDUX

In this section, I set out what I take to be a better conception of autonomy. This serves two ends. First, it offers a charitable addition to Elster's own account. If, as I suggested in §2, we should interpret him as saying that 'autonomy' refers to whatever substantive standard against which our preferences should be assessed, then what follows offers such a standard which is consistent with the schematic theory that he has laid out. Secondly, supplementing Elster's view with the following theory of autonomy allows us to address the two problems for Elster that I have been discussing: first, by giving a clear account of the badness of adaptive preference formation, and secondly by showing how we can reconstruct from the apparently diverse proposals in §3 a unified and principled distinction can be drawn between that phenomenon and character planning.

The conception of autonomy I propose is broadly the same as what Joseph Raz has in mind when he describes autonomy as an 'ideal of self-creation', and speaks of an agent as 'part author of his life'.[25] These metaphors are evocative, but somewhat vague. Elsewhere, I have suggested the following formulation:

> Autonomy consists in deciding for oneself what is valuable, and living one's life in accordance with that decision.[26]

This seems to me the most defensible of the various conceptions of autonomy in the intellectual marketplace. As it stands, the definition raises a number of questions, though. Some are hermeneutical (how far would Raz say that this identifies the same idea as his?). Others are justificatory (should we think that autonomy, so conceived, is valuable?). I shall address neither set of questions seriously here, though I take my conclusions in this paper to be relevant to both. The success of my conception in dealing with the problems I have identified in Elster should serve both as a weak recommendation for conceiving of autonomy in the way that I do, and as a defence against the arguments in §§2-3 insofar as they might be mobilised as objections to a political theory committed to autonomy as an ideal. A third set of questions concern the details

[25] J. Raz, *The Morality of Freedom* (Oxford, 1986): p. 370. Similar notions of autonomy can be found in T. Hurka, *Perfectionism* (Oxford, 1993) at p. 148; and S. Wall *Liberalism, Perfectionism and Restraint* (Cambridge, 1998), at p. 128.

[26] B. Colburn, *Autonomy and Liberalism* (New York, 2010): p. 19.

of my conception of autonomy: for example, what does 'deciding for oneself' mean? I address *these* questions in the remainder of this section.

Before doing so, I note that a conception of autonomy like this can plug the gap I noted at the conclusion of the previous section, by motivating the connection between higher-order reflection and autonomy. Different theorists might construe this connection in different ways. Perhaps high-order reflection might be deemed both necessary and sufficient for self-authorship, in which case the two conceptions of autonomy end up extensionally equivalent, and the reference to self-authorship merely serves to motivate taking this as an ideal. On the other hand, one might think higher-order reflection merely necessary but *not* sufficient, or (as the tenor of my discussion above perhaps implies) neither, though it is generally supportive of self-authorship. On those views, autonomy on Dworkin's conception will turn out to be constitutively, or instrumentally, or heuristically valuable with respect to autonomy on my conception. In any case, though, such dependence would add weight to the thought that mine is a more fundamental ideal than those of rival theorists of autonomy.

To recap: autonomy consists in deciding for oneself what is valuable, and living one's life in accordance with that decision. That has two principal components. The latter deals with success in pursuing one's aims, and is not relevant to our present discussion. The former concerns the conditions those aims must meet if their pursuit is to count towards our autonomy. Now, the word 'decide' is ambiguous in ordinary usage. It can refer to a choice by an agent, or to some sort of epistemic judgment. These are usually distinguished by the uses of the locutions 'decided to' and 'decided that', respectively. I do not intend to presuppose either usage when I say that autonomy involves people 'deciding for themselves' what is valuable. Indeed, the double meaning seems appropriate: some individuals will choose to pursue some project and thereby make its fulfilment valuable, and some individuals will reflect and decide that such-and-such an end is valuable. The crucial thing is that an agent decides for herself (in the sense relevant to autonomy) to the extent that the following two conditions hold:

- **Endorsement** – she has a disposition such that if she reflects (or were to reflect) upon what putative values she ought to pursue in her life, she judges (or would judge) of some such things that they are valuable.
- **Independence** – She is in a state where her reflection is, or would be if it took place, free from factors which limit the extent to which we can say that she is deciding for herself.

The Endorsement Condition requires the presence only of a disposition. So, an agent can satisfy it without necessarily going through the process of consciously reflecting upon her values: it may just be that *were* she so to reflect, in her present circumstances, she would come to the judgement described. This focus on a disposition is a way of making concrete the attractive notion that what matters is not the act of occurrent reflection *itself*, but rather the relationship that such reflection reveals between an agent and the commitments which shape her life. One can endorse values either explicitly or implicitly, but that crucial relationship obtains in both cases. So, on my theory, one need not have consciously reflected upon whether one really takes a given thing to be valuable to be autonomous in its pursuit: instead, one's behaviour might indicate a tacit endorsement of that value. So, consider someone who is a talented geneticist and pianist, and who eventually chooses to pursue the cure for cancer rather than the world of concert performance. It may well be that she never consciously weighs up two different putative values –

'curing cancer' and 'producing great music' – and makes an explicit judgment about which one she believes she should pursue. Nevertheless, we might think that her pursuing the cure for cancer is an implicit endorsement of curing cancer as a valuable pursuit: she was aware of what alternatives she had, and might have explained, if we asked her to, why she took that course rather than pursuing the musical life instead. Phrasing the Endorsement Condition in terms of dispositions allows us to say that an implicit endorsement like this also counts as deciding for oneself what is valuable.

The Independence Condition is a good deal vaguer than its companion: it requires that an agent be in a position such that her reflection is (if it takes place) our would be (if it were to take place) free from factors which undermine the extent to which we can say that she is deciding *for herself*. This captures something important, but gives us little help if we want a general account of when someone's independence is undermined. One way we might try to make things clearer is to return to Dworkin, who also insists on the importance of procedural independence (as he puts it). Dworkin admits that he can give no general account himself, but he does at least give a succinct explanation of what sort of account is needed, which can serve as our starting point:

> Spelling out the condition of procedural independence involves distinguishing those ways of influencing people's reflective and critical faculties which subvert them from those which promote and improve them. It involves distinguish those influences such as hypnotic suggestion, manipulative coercive persuasion, subliminal influence, and so forth, and doing so in a non ad hoc fashion.[27]

As Dworkin notes, independence in the relevant sense does not mean the absence of *any* influences on our decisions about what is valuable. Only those who are hostile to a concern for autonomy would set up such a straw man. Rather, when we say that someone's commitments are independent, we mean that they are free of a *certain sort* of influence, which is instantiated in the intuitive instances listed above. The challenge that Dworkin lays down is therefore to identify this baleful influence.

5. INDEPENDENCE AND COVERT INFLUENCE

In what follows, I give a partial answer to Dworkin's challenge, by proposing a necessary condition for independence, which centres on the notion of *covert influence*. Someone's commitments (or values, or judgements, or preferences – for present purposes it doesn't really matter which) are covertly influenced when the explanation for those commitments is something that is necessarily hidden from them, in the sense that it would *not* be the explanation for their commitment if it weren't hidden. And when they are covertly influenced, they lack independence.[28]

---

[27] Dworkin, *Theory and Practice of Autonomy*: p. 18.

[28] There is, of course, a further question of whether the lack of covert influence is not merely necessary, but also sufficient for independence. Since an answer to that question is not needed for my purposes here, I do not seek to address it.

This, I suggest, is the shared factor which is at work in the various cases that Dworkin lists. Hypnotic suggestion and subliminal influence both work necessarily through bypassing someone's conscious deliberative processes. So, to use an example borrowed from Roger Crisp, imagine a case of subliminal advertising by a cinema. Single-frame adverts for ice-cream are flashed on screen during the showing of a film, as a result of which people in the audience form a desire to eat ice-cream during the interval. In such cases, the explanation for their preference for ice-cream is covert: they cannot be aware of it. It seems likely that from their point of view this preference is based on a proper appreciation of the virtues of ice-cream. At any rate, if their preference is genuine they *won't* think that the only reason they have for it is that single-frame images of ice-cream have been interspersed with the film that they have been watching. Nevertheless, as detached observers we can see that this is exactly what has happened. They entered the screening with no preference for ice-cream, and left with a marked preference and intention to buy, and the reason for this is the subliminal technique that has been applied to them.[29] So, the explanation for their preference is necessarily hidden from them – if it weren't, it wouldn't be the right explanation.

As Crisp points out, that is not to say that the technique itself is necessarily hidden. One can be informed that one has been the subject of subliminal messaging. One can even be so informed without that causing the artificially induced desires to lapse. The point is just that *when* someone is made aware of that, the explanation for their preference must change. We no longer say just that they desire ice-cream because the cinema management induced the desire in them. Depending on their reaction, we would say either that they realize that the desire was induced but can adduce other independent reasons for their eating ice-cream being something they want to do, or that they repudiate that desire (in which case, as Crisp notes, the ostensible innocent desire for ice-cream has become the sort of unwanted craving which is a paradigmatic threat to autonomy).[30] The point is that subliminal messaging is covert *insofar as it is the explanation* of our sincerely held preferences.

Subliminal advertising and hypnotic suggestion are somewhat spectacular example cases. Crisp discusses various more mundane techniques used by advertisers which he thinks are damaging to a consumer's autonomy. In general, the effective techniques of persuasive advertising are effective precisely because they play on the subconscious, and therefore create distance between the explanation that a consumer adduces for their preference and the explanation that an impartial observer would be inclined to give. 'When I buy Pongo Peach [cosmetics],' Crisp says, 'I may claim that I want to look good. In reality, I buy it owing to the link made by persuasive advertising between my unconscious desire for adventure and the cosmetic in question'.[31] Moreover, we can see that the mechanism is not merely hidden, but covert: the mechanism *must* be hidden because it wouldn't work otherwise. To make the point, Crisp asks: would you buy Pongo Peach products if they advertised it by saying 'Do you have a sense of adventure? Then use this brand of cosmetics'. When the attempt to link Pongo Peach with the subconscious desire for adventure is made explicit, it is also made risible, and hence uneffective.[32]

---

[29] R. Crisp, 'Persuasive Advertising, Autonomy, and the Creation of Desire', *Journal of Business Ethics* 6 (1987): 413-18.

[30] Crisp, 'Persuasive Advertising': 414-5.

[31] Crisp, 'Persuasive Advertising': 415.

[32] Crisp, 'Persuasive Advertising': 416. We can assume that the unconscious link is indeed risible, and hence won't stand up to scrutiny.

I don't mean to argue here in favour of Crisp's claim that such mechanisms are endemic in advertising (though as it happens I think it's true). The point here is just that Crisp identifies the right problem: the influence involved – and hence the explanation for agents' commitments – is *covert*, in the sense described above.

Covert influence on an agent's commitments is bad for autonomy, because it undermines the extent to which we can say she herself is deciding on what is valuable. To illustrate the point, consider the difference between first- and third-person explanations for a person's commitments. Usually, the former will feature in the latter. If someone asks me why I am committed to playing a musical instrument, I might say something like 'Because I devoted myself to learning the harp several years ago, and it is important to me to fulfil that ambition', or 'Because playing the harp well is valuable'. In most cases, someone else trying to explain my commitment will echo these answers: 'It is because he wants to fulfil his ambition to succeed in his chosen hobby', or 'It is because he believes that playing the harp well is valuable'. And that is as it should be: when thinking about why someone has the commitments they do, their own perspective on what is valuable and their motivations has some sort of authority. By contrast, there are cases in which the first-person explanation features in none of the third-person explanations because it is irrelevant. If someone is brainwashed into joining a cult, the third-person explanation for her commitment will *not* take at face value her rapturous account of seeing the light. If she is subliminally influenced into wanting ice-cream, then the third-person explanation will disregard her attempts at rationalising her sudden longing for raspberry ripple. In such cases, the 'real' reason for her commitments is opaque from her first-person point of view. Because something else (about which she cannot know) explains her commitments, we can't say that she is deciding for herself. So, her autonomy is compromised, because she fails the Independence Condition in respect of these commitments.

To recap: the proposal is that autonomy is undermined when our commitments have covert explanations. Focussing on *covert* explanations, as opposed to ones of which we're merely unaware or unconscious, is important for three reasons.

The first is that it best captures the intuitive thought that the problem is not just that the explanation for a commitment happens not to have occurred to an agent, but rather that it *could not* occur to them. The second is that it echoes the reasons given above for phrasing the Endorsement Condition in terms of a disposition to endorse, rather than requiring occurrent reflection. As I said there, requiring occurrent reflection would lead to an implausible and narrow conception of autonomy, fetishising rational reflection (rather than regarding it as a useful indication of what is actually important) and excluding various obviously autonomous lives, like that of the devoted but unreflective cancer scientist. *If* the Independence Condition required that we be aware of all explanations for our commitments, then it would have a similar effect. Picking out only covertly explained commitments as problematic avoids this, and hence harmonizes with the reasons given above for preferring to think about autonomy in the way that I proposed.

The third reason for focussing on covert influence is that it allows me to sidestep the criticisms I levelled against Elster in §3B. There, I noted that identifying the distinction between adaptive preference formation and character planning would have various implausible consequences, chiefly because it would have to condemn as problematic any preference not conscious formed. Insisting that the problem is not with unconsciousness per se but with *necessary* unconsciousness means that I am not vulnerable to the same problem. And it also allows us to pay due heed to the impression that – the aforementioned criticisms notwithstanding –

Elster had put his finger on something important when he noted that there's something troubling about our commitments being formed behind our backs.

6. COVERT INFLUENCE AND ADAPTIVE PREFERENCE FORMATION

In the previous section, I set out a conception of autonomy on which the notion of covert influence is central. If our commitments have covert explanations – that is, explanations of which we necessarily are not aware – then those commitments lack independence, and lack of independence undermines autonomy. It remains to show how this way of thinking about autonomy allows us to offer solutions to the two problems I set out with Elster's theory: how to account for the badness of adaptive preference formation, and how to distinguish between that phenomenon and character planning.

First, recall that Elster's attempt to explain the badness of adaptive preference formation failed. To motivate the thought that there is something wrong with adaptive preference formation, we needed a prior notion of autonomy. Not only did Elster fail to give us an explicit definition, the implicit definition – that autonomy is what is left after the mechanisms like adaptive preference formation have been eliminated – made the account circular.

The account of autonomy and independence developed in the previous two sections allows us a charitable modification of Elster's theory. Let us understand autonomy, as I have suggested, as consisting in an agent deciding for herself what is valuable, and living her life in accordance with that decision. This means that anything which violates the Independence Condition undermines autonomy, and (so long as we think autonomy valuable) is bad for that reason. Adaptive preference formation, however, is a paradigm case in which the explanation for our preferences is covert. The fox explains his preference by saying that the grapes are sour – but we know better, and explain it by referring to his unconscious downgrading of the inaccessible option. That explanation is covert, for the fox couldn't be aware of it and it still be the right explanation of his preference change. For one thing, it would no longer be unconscious. More importantly, even if we don't want to take adaptive preference formation to unconscious by definition, the fox could not explain his preferences on the basis of a belief that the grapes are sour if he is aware that their inaccessibility is the only reason he has that belief.

So, supplementing Elster's account of adaptive preference formation solves the first problem, by giving a clear reason why adaptive preference formation is bad.

The conception of autonomy I have laid out also gives us a principled distinction between character planning and adaptive preference formation. Both are ways of dealing with 'a state of tension between what you can do and what you might like to do'.[33] However, they differ in that the preferences we end up with admit of different explanations. As Elster describes it, character planning is never covert: it is always a conscious procedure of 'trying to shape one's wants so as to coincide exactly with – or differ optimally from – one's possibilities'.[34] Adaptive preference formation, by contrast, always is. This is a clean distinction – and it also shows why Elster is right that character planning is not problematic for the same reasons as adaptive

[33] Elster, *Sour Grapes*, p. 117.

[34] Elster, *Sour Grapes*, p. 118. Interestingly, the *means* of character planning employed might be covert, even if the crucially significant decision to engage in the process is not. So, for example, if I fail at overt character planning, I might decide to put myself in the hands of someone who is a master at covert preference change, in the hopes that their covert techniques might be successful. My thanks to an anonymous referee for the example.

preference formation, despite the structural features that the two phenomena share. So long as the Endorsement Condition is satisfied – that is, so long as one has the disposition on reflection favourably to assess the character ideal in light of which the planning takes place – then character planning can be actively supportive of an individual's autonomy, though it need not be.[35]

This proposal for drawing the distinction between adaptive preference formation and character planning is coextensive with none of Elster's three attempts, as catalogued in §3 above. However, it does also show that there was a grain of truth in each attempt.

Proposal *A* was that adaptive preference formation is distinctive because it is a purely causal process. My proposal is orthogonal to this. A covert explanation for a preference is not *ipso facto* a causal explanation, nor is a non-covert explanation *ipso facto* non-causal. Whether or not we think there can be non-casual covert explanations or non-covert causal ones will depend on our conclusions in other bits of philosophy, but for present purposes we need only observe that the question is irrelevant to distinguishing adaptive preferences and character planning.

Elster might, of course, say that he was using the term 'causal' more loosely than I have given him credit for, and that what he meant was just that adaptive preference formation is a process whereby an agent's preferences can be explained without referring to the explanation which they themselves would be inclined to give of their action – it eliminates the important first-person authority which I referred to in §5. If something like *that* is what Elster meant, then he was getting towards the truth – but then he has reason to accept my conception of autonomy and the account of the distinction which flows from it, as the most coherent way of paying heed to the intuition he was trying to capture.

Proposal *B* was that the distinction is between an unconscious process (adaptive preference formation) and a conscious one (character planning). My proposal differs from this, for although covert influence must be unconscious, unconscious processes need not be covert. So, Proposal *B* draws the line in the wrong place, and incorrectly counts some innocent processes of preference formation as adaptive.

Once again, though, Elster was correct to identify as crucial the idea that adaptive preference formation is unconscious. My proposed distinction could therefore just be read as making this intuitive idea sharper by stating explicitly that the problem with adaptive preference formation is that it *has* to be unconscious, not merely that it happens to be so.

Proposal *C* was to construe the distinction in light of Dworkin's account of autonomy – character planning consists in lower-order preferences being shaped by higher-order ones, and adaptive preference formation consists in them being shaped by 'drives'. My proposed distinction is orthogonal to this too. There might instances of higher-order shaping of preferences which nevertheless count as adaptive preference formation on my view because the higher-order preferences might themselves be ones which have covert explanations. And there need be nothing covert about the process whereby a strong first-order desire affects other first-order preferences: the example I gave in §3C of someone who (mindful that she is perpetually and powerfully hungry) consciously cultivates preferences for cheap victuals seems to be an

[35] For instructive discussion of this, see Zimmerman ('Making do': 35-7, and 'Sour grapes': 225-26), who worries that on Elster's view we can't distinguish character planning from the much more troubling phenomenon of *self-abnegation*, whereby an agent consciously seeks to eliminate desires that lead to unhappiness due to dramatically curtailed option-sets.

uncontroversial case of character planning.[36] So, appealing to Dworkin's conception of autonomy leads to Proposal *C* identifying the wrong distinction.

Elster's basic tactic is sound, though. If I am right, the distinction between the phenomena does indeed piggyback on an account of autonomy. Given my criticisms of Dworkin's view in §3C are persuasive, someone who wants to make use of an ideal of autonomy has reason to shift to my conception – and these reasons are internal to a concern for autonomy, irrespective of my position's ability to solve Elster's two problems. So, Elster himself could modify his account of adaptive preference formation along the lines I've suggested, without being vulnerable to the charge of *ad hoc* squirming.

CONCLUSION

I started this paper by setting out a pair of problems with Jon Elster's influential account of adaptive preference formation: he gives neither a reason to think the phenomenon a bad thing, nor a clear and principled way of distinguishing it from the less malign process of character formation. In both cases, the problem turned out to be with Elster's conception of autonomy – either because he gave only a circular and therefore unilluminating definition of the ideal, or because he relied on a conception of autonomy (namely Gerald Dworkin's) which could not do the work he needed it to do. My proposal has been that we explicitly define autonomy as consisting in an agent deciding for herself what is valuable, and living her life in accordance with that decision. An important threat to autonomy, so conceived, is loss of independence – that is, a diminution of the extent to which we can say that an agent decides for herself. I suggested that this happens when she is covertly influenced – that is, when the explanations for her commitments are necessarily hidden from her. Adopting this view is attractive on its own merits. And, when coupled to Elster's account, it solves the two problems mentioned above. Adaptive preference formation is bad because it is always covert, whereas character formation is always non-covert.

In case I be thought guilty of painting an altogether too rosy picture, I conclude a caveat. My proposed rescue for Elster only works if my conception of autonomy is defensible, and I have offered only small and indirect argument for that claim in this paper. So, accepting my proposal is not costless. Someone who insists on understanding autonomy some other way – in Dworkin's sense, for example, or perhaps as a more overtly Kantian conception of self-legislation – will not find my account persuasive. Neither will someone who thinks that autonomy is not in fact an important ideal.

These possible sources of disagreement should not worry us too much, though. I have already said something to the first critic, by commenting on some reasons to think that other conceptions of autonomy are problematic. And the second critic seems to bear a burden of proof, in that she must show why – if we don't really care about autonomy at all – we should think adaptive preference formation problematic in the first place.[37]

---

[36] Of course, we might think there's still something wrong with her situation, from the point of view of autonomy or otherwise.

[37] My thanks to Harry Adamson, Daniel Elstein, Lorna Finlayson, Hallvard Lillehammer and Serena Olsaretti for discussion on arguments in this paper.