

## *Autonomy and Addiction*

NEIL LEVY

Centre for Applied Philosophy and Public Ethics

University of Melbourne

Parkville, 3010

Australia

and

James Martin 21st Century School

Oxford University

Oxford OX1 1PT

United Kingdom

Whatever its implications for the other features of human agency at its best — for moral responsibility, reasons-responsiveness, self-realization, flourishing, and so on — addiction is universally recognized as impairing autonomy. But philosophers have frequently misunderstood the nature of addiction, and therefore have not adequately explained the manner in which it impairs autonomy. Once we recognize that addiction is not incompatible with choice or volition, it becomes clear that none of the standard accounts of autonomy can satisfactorily explain the way in which it undermines fully autonomous agency. In order to understand to what extent and in what ways the addicted are autonomy-impaired, we need to understand autonomy as consisting, essentially, in the exercise of the capacity for *extended agency*. It is because addiction undermines extended agency, so that addicts are not able to integrate their lives and pursue a single conception of the good, that it impairs autonomy.

### **I Accounts of Autonomy**

Available accounts of autonomy fall into two broad classes: *procedural* and *substantive* (Mackenzie and Stoljar, 2000). Substantive accounts place restrictions on the kinds of preferences compatible with autonomy, whereas procedural accounts are neutral with respect to the content of preferences. Substantive and procedural accounts further divide into *structural* and *historical* procedural accounts, on the one hand, and *strong*

and *weak* substantive accounts, on the other; there are also accounts that combine substantive and procedural elements. In what follows, I shall not attempt to address every theory in all its variants. Instead, I shall briefly sketch the main lines of some of the better known, with the aim of showing how and why they fail to give the right result when they are applied to at least some cases of addiction.

First, then, a lightning tour of some of the better known accounts of autonomy, beginning with the procedural accounts and moving through to the substantive. The best known account of autonomy is the structural theory associated with Harry Frankfurt (1988). On this view, roughly, someone counts as autonomous if she identifies with her effective first-order desire, where identification is cashed out in terms of a special higher-order desire that that first-order desire be her will. Autonomy is thus a question of the structure of the agent's hierarchy of desires.

This structural account of autonomy is subject to what some see as a devastating counterexample, from cases of manipulation (Slote, 1986; Fischer and Ravizza, 1998). An agent whose preferences are systematically manipulated by a neuroscientist of whose presence she is not aware fails, intuitively at least, to count as autonomous. Yet such an agent might identify with her effective first-order desires. Indeed, there seems no reason, in principle, why her identification might not itself be subject to manipulation. For this reason (amongst others), some philosophers have urged that we place *historical* constraints upon our conception of autonomy.

Important historical accounts of autonomy have been developed by Gerald Dworkin (1988) and by John Christman (1991). Christman, for instance, argues that for a preference to be autonomous, it must pass a historical test: the agent must endorse not only her preference, but also (actually or counterfactually) the process by which she came to acquire it. However, historical accounts themselves seem vulnerable to an objection of much the same kind as that which motivated proponents to reject the Frankfurtian view. It seems that we can as easily be manipulated into endorsing a process of preference acquisition as into endorsing a preference.

Indeed, I can be manipulated into endorsing the very act of manipulation; some philosophers have argued that we should see the preferences of women in some sexist cultures as manipulated in just this way. In fact, there seem to be plenty of real-life cases in which agents who are the victims of extremely unjust societies not only endorse their socialized preferences, but also continue to endorse them even after they become aware of the manner in which they acquired them. For this reason, some philosophers have argued that an adequate account of autonomy must be substantive; that is, it must place constraints on the content of preferences that are to count as autonomous. For instance, Paul Benson (1994) argues that only agents who regard themselves as competent to answer for their conduct in the light of normative demands can count as autonomous.

How should we go about deciding which of these competing accounts of autonomy is the best? One way we might proceed is by asking which best captures our everyday and philosophical uses of the word. I think that this is an unpromising approach. 'Autonomy' is used in too many different ways for this kind of approach to work (Arpaly, 2004). Sometimes it is used as a synonym for freedom, sometimes it is distinguished from it; sometimes it is used in a normatively laden way, such that only lives that are flourishing can count as autonomous, sometimes it is used in a more neutral manner. If we are to make progress, we need to restrict the range of meanings we attribute to the word. On the other hand, we cannot settle debates concerning the analysis of a concept by fiat: we are not allowed to use words however we please. The analysandum must bear some significant relation to the everyday concept, if we are to avoid the danger of talking past each other.

One reason for the relative lack of progress in convergence on a single sense of the word, I suggest, is that under the rubric 'autonomy' there are several, linked but different, concepts at issue. What sense of autonomy ought to concern us here? That is, what kind of autonomy does addiction impair? In what follows, I shall restrict myself to what Christman (2003) calls *basic* autonomy: 'the minimal status of being responsible, independent and able to speak for oneself.' Basic autonomy contrasts with *ideal* autonomy, which is the state of maximal authenticity to which many aspire, but few achieve. Basic autonomy, I claim, is a procedural notion. I therefore propose the following, deliberately rough and admittedly somewhat stipulative, definition of autonomy, as I shall use it here. Autonomy is *self-government*; the autonomous individual is responsible for her actions because they express her will, and they express her will because the dispositions which they put into effect are hers in some important sense.

We might usefully distinguish *self-government* from *self-determination*. Self-government is a matter of having and acting upon the preferences one wants to have; self-determination is something much more demanding. To be self-determining is to be the *source* of the preferences upon which one acts; to have made them one's own in some deep sense. Self-determination might be cashed out in incompatibilist terms; one might be self-determining if one is *ultimately* responsible for one's preferences and their structure (Kane, 1996). Or it may be understood in terms of *authenticity*: one is self-determining if one's preferences are one's own in some deep sense. But self-government requires much less: neither ultimate responsibility nor authenticity; it is, to that extent, a much shallower notion. It is self-government, I shall claim, that is sufficient for basic autonomy; self-determination seems instead to map onto the notion of ideal autonomy.

This definition may seem to beg the question against substantive conceptions of autonomy. To some extent, it is designed to do so, on the grounds that basic autonomy just *is* a procedural concept. It is reasonable to think that it is impossible to elaborate a merely procedural conception of human *flourishing*, or of *ideal* autonomy, but it does not follow that an analysis of *basic* autonomy cannot be content-neutral. If autonomy is self-government, then (absent further argument) it seems that an autonomous individual will not necessarily live a flourishing life. She need not be happy, or have largely true moral or nonmoral beliefs. Nor (more contentiously) need she enjoy liberty, at least not to its fullest extent. The autonomy debate is not the free will debate, nor is it the moral responsibility debate (though basic autonomy is closely connected to these notions, it is not identical to them). Self-government is a procedural notion, and it is at least *prima facie* a near synonym for autonomy in its basic sense.

Nevertheless, friends of more substantive conceptions of autonomy needn't think that my account has nothing to offer them. We ought not to build substantive conditions into the *definition* of autonomy, but we might discover that autonomy, procedurally defined, has substantive preconditions. I shall argue that this is in fact the case; the account on offer therefore implies (relatively weak) substantive elements. For this reason, the conception of autonomy as self-government is able to give the right result in cases that defenders of substantive conceptions have, rightly, taken to be examples of autonomy impairment.

These brief remarks do not constitute, and are not intended to constitute, a defence of the view that autonomy should be understood as procedural self-government. Rather than defend this view directly, I am content to allow its virtues to emerge from the way in which it, and it alone, handles the autonomy-impairment characteristic of addiction.

## II Addiction

The debate between the proponents of various accounts of autonomy has for the most part focused, on the one hand, on more or less fantastic examples of manipulation, and, on the other, on all too real cases of oppressive socialization. Addiction has played a relatively small role, at least until recently.<sup>1</sup> The reason why it has played so small a part in this

---

1 The past few years have seen a flowering of an important philosophical literature on addiction (for a review of some of the highlights of this literature, see Yaffe, 2002). This literature has begun — though it is far from completing — the necessary task of correcting the mistaken conception of addiction as compulsion with which philosophers have typically worked.

debate, I suspect, is that it is felt to be too easily dealt with. So devastating are its effects on the will that addicts do not even get into the autonomy ball-park. Addiction does not merely impair autonomy, it is widely felt; instead its effects go much deeper, destroying agency itself.

On this conception of addiction, addicts literally can't help themselves. This is a view that found its most eloquent expression over a century ago, in the writing of William James:

The craving for a drink in real dipsomaniacs, or for opium or chloral in those subjugated, is of a strength of which normal persons can form no conception. 'Were a keg of rum in one corner of a room and were a cannon constantly discharging balls between me and it, I could not refrain from passing before that cannon in order to get the rum'; 'If a bottle of brandy stood at one hand and the pit of hell yawned at the other, and I were convinced that I should be pushed in as sure as I took one glass, I could not refrain': such statements abound in dipsomaniacs' mouths. (James, 1890, 543)

This is a view of addiction which is still widely held, not only by philosophers, but also by bioethicists and psychologists. For Louis Charland, for instance, 'the brain of a heroin addict has almost literally been hijacked by the drug' (Charland, 2002, 43). For Carl Elliott, the addict 'is no longer in full control of herself. She must go where her addiction leads her, because the addiction holds the leash' (Elliott, 2002, 48). For Alan Leshner, the initially voluntary behavior of drug-taking gradually transforms into 'involuntary drug taking, ultimately to the point that the behavior is driven by a compulsive craving for the drug' (Leshner, 1999, 1315). Even the *Diagnostic and Statistical Manual* of the American Psychiatric Association holds that addiction 'usually' involves 'compulsive drug taking behavior.'

However, if 'compulsion' is taken literally, this is false. That is, if by a compulsive force we mean one that bypasses the agent's will entirely, or one that cannot be resisted by her, then it is false that addictive desires are compulsive. The addict is not carried away by her desires in the way in which, in Aristotle's illustration of non-voluntariness, a man is carried across the road by the wind. The point is not that there is no such thing as compulsion by forces internal to the agent. The point is that, whether or not there are compulsive psychological forces, addictive desires are not among them.

Indeed, if addictive desires were compulsive, it is difficult to see how addicts could give up voluntarily. But they do, in their thousands, largely without assistance from others (Sobell, Ellingstad, and Sobell, 2000). There is plenty of direct evidence, in any case, that addicts exercise some degree of control over their consumption behavior. Both within and outside the laboratory, consumption is price sensitive, in a manner that would be surprising if addictive desires were compulsive (Elster, 1999;

Neale, 2002); it remains price sensitive even after an initial dose of the drug of addiction (Fingarette, 1988, 36-42). It is widely believed that either the craving for the drug, or the fear of withdrawal, is so powerful as to overwhelm the volitional resources of addicts. But the typical addict goes through withdrawal several, perhaps many, times. Indeed, some deliberately abstain for prolonged periods in order to lower their tolerance for the drug, and thereby decrease the dose they will need to achieve the high they want (Ainslie, 2000, 82). Addicts do indeed experience cravings — more intensely for some drugs than for others — and withdrawal is indeed an unpleasant experience (though once again the extent to which this is so varies from drug to drug; cocaine addiction seems to be almost entirely a matter of craving and not withdrawal). But rarely or never are these forces, singly or combined, sufficient to move the addict against her will.

What, then, explains the autonomy impairment characteristic of addiction? In what sense can they truthfully claim to act against their wills? Some philosophers have suggested that the primary impairment associated with addiction is coercion: addicts act against their wills in order to avoid the painful experience of withdrawal (Watson, 1999). But not all addictions are associated with withdrawal, and addicts seem to remain autonomy-impaired even *after* they have undergone detoxification, so that withdrawal no longer threatens them. Even absent the pull of craving and the push of withdrawal, addicts continue to consume their drugs. That is, even after they have thrown off the chemical addiction, which produces a characteristic cycle of craving, consumption, satiation and craving, addicts are likely to relapse (between 40% and 60% of addicts return to using after apparently successful treatment [Bonnie, 2001]). Addicts are not driven to use by drug-induced alteration in neuropsychology, though the adaptation to drug use that occurs in the dopamine system in the brain is real enough.

None of this is intended to deny that withdrawal may not have a coercive effect upon addicts. I simply want to suggest that, whatever the truth about coercion, it is not the whole story. There is an impairment of autonomy characteristic of addiction, in addition to whatever coercive effects withdrawal may have.

So why do addicts consume their drugs? The short and only somewhat misleading answer is that they take drugs because they want to. Indeed, there is a very real sense in which they choose to take their drug. Only the hypothesis that they want to consume can explain why they inject themselves, why they engage in instrumentally rational actions designed to procure their drug or the money they need to buy it, or why they are likely to readdict themselves after withdrawal. It is not compulsion, or coercion; it is, in some sense, volition.

In the face of this kind of conclusion, it is tempting to become an addiction sceptic: that is, to conclude that addiction no more undermines autonomy than does, say, a desire for strawberries (Foddy and Savulescu, 2006). But as all of us who have ever struggled with an addiction — whether to caffeine, tobacco or to heroin — know, that is far too hasty. Addicts say that they use against their will, and there does seem to be some sense in which this is true. After all, not only is there the phenomenological evidence, to which many of us can attest, that breaking an addiction is difficult, there is also the evidence that comes from the fact that all too often addicts slowly destroy their lives and the lives of those close to them. They engage in illegal, dangerous or degrading activities in order to procure their drug, they lose their jobs, their partners and their homes. If it was *purely* a matter of autonomous choice, we should not expect their lives to spiral out of control so dramatically. Addicts frequently say that they consume against their wills; I shall argue that there is a sense in which this is true.

### III Addiction and the Oscillation of Preferences

I suggest that we reconcile the evidence that addicts are autonomy-impaired and the discovery that they take their drug because they want to by understanding unwanted addiction as characterized by an oscillation in the preferences of the addict. Most of the time, the addict sincerely disavows her addiction and wishes to be rid of it. But she regularly changes her mind; when she does, she genuinely prefers consumption to abstention.

George Ainslie's work on time inconsistency of preferences provides a useful perspective on the kind of oscillation here (Ainslie, 2001). Economists and psychologists have generally supposed that we discount future goods exponentially. Exponential discounting explains some kinds of inconsistency, but, arguably, not the kind characteristic of the addict. Suppose the addict discounts exponentially both a drug-free existence and the immediate pleasure of consumption. In that case, the closer in time her opportunity for consumption, the higher her estimate of its value. But, if her discount curve is exponential, then she discounts the value of a drug-free life just as much as she discounts the value of consumption. If, just before she consumes, she prefers consumption to abstention, and her discount curve is exponential, then at *any* time prior to consumption she will prefer consumption to abstention. If, therefore, she claims that she acts against her will, she is lying or self-deceived; she chooses what she genuinely prefers.

Exponential discounting can explain regret in one-shot games (as it were): it can explain why an agent might regret choosing *X* over *Y*, where

choosing  $X$  at time  $t$  precludes choosing  $Y$  at time  $t1$  (and  $t1$  is later than  $t$ ). But it cannot explain the oscillation of preferences characteristic of the addict. However, if our discount curves are hyperbolic — that is, highly bowed — our discount curves can cross. The closer in time to us the good, the steeper the curve, and the more likely it is that it will cross other curves, which express our valuation of a good further in the future. As a result, our estimate of the value of future goods can be inconsistent: one and the same agent can prefer future good  $X$  to future good  $Y$  at time  $t$ , and  $Y$  to  $X$  at  $t1$ . We therefore get time inconsistency of behavior. Suppose that  $X$  and  $Y$  are mutually exclusive goods (for instance sticking to my diet and eating sticky date pudding). At  $t$  I prefer  $X$  to  $Y$ , but as the time at which  $Y$  is accessible approaches, the steepness of my discount curve increases. The value of  $Y$  outweighs the value of  $X$  for me at  $t1$ , when I make my decision. Sooner or later, I regret my decision, and revert to my previous weighing of  $X$  and  $Y$ . But, unless I take steps to avoid the cycle repeating itself, I am destined to reverse my weighting of the two goods once again.

Suppose that Ainslie's theory, or something like it, is correct; in what way are addicts autonomy-impaired as a consequence of their addiction? A hierarchical account of autonomy, like Frankfurt's, will explain impairment in terms of a conflict in the conative hierarchy of the agent: she acts upon a desire she disendorses, and therefore acts against her own will.<sup>2</sup> Having a desire we disendorse is a common enough experience: we have resolved to stick to a diet, and are dismayed to find ourselves salivating when the dessert trolley arrives. Perhaps people are sometimes even moved all the way to action by a desire that, all things considered, they reject. But addicts are not like that, not, at least, in all cases (and probably not even in most). Instead, addicts change their minds: the opportunity for consumption arises, or the cravings begin, and the pleasures of the drug begin to weigh more heavily with them than the goods achievable through abstaining. Perhaps the focus on consumption 'crowds out' other considerations, so that the value of other options are no longer keenly appreciated by her; perhaps the coercive effects of withdrawal or the attractive force of craving lead her to value consumption more highly than previously. She is not moved by a desire that is alien to her; instead, she is moved by what seem to her, at the time of action, to be good reasons.

---

2 Frankfurt would reject this claim, since he identifies the agent's will with her effective first-order desire. He would prefer to say that the agent is unwilling because she does not act upon the will she wants. I think my way of phrasing the claim better captures the usual way in which we use the word — but nothing substantive seems to turn on this question.

In Richard Holton's useful phrase, the experience of craving induces 'judgement shift' in the addict (Holton, 2004).

A proponent of the hierarchical account might argue that the unwilling addict will nevertheless experience some kind of motivational conflict, even as she decides to consume. She may continue, to some extent at least, to value abstaining. Ainslie's preferred solution to weakness of the will, the adoption of what he calls 'personal rules,' depends upon this being the case. An agent adopts a personal rule when she bunches the rewards of future abstention together, seeing her current decision to abstain as setting a precedent for her future behavior; for such a rule to work, she must value future abstention even as she is tempted to consume. But, first, this need not be the case: an addict can count as unwilling, in a sense we shall clarify, even if at the time of consumption she prefers consumption now *and* in the future to abstention (unlike Ainslie's addict, who prefers consumption now, but future abstention). And, second, even if the addict does experience the kind of conflict in question, this fact alone does not amount to autonomy-impairment. Suppose she abstains, now and on every future occasion upon which drug-taking is an issue for her. Will we say that she is autonomy-impaired if, on each occasion, she has a lively appreciation of the value of consumption? Recall the Alcoholics Anonymous slogan: 'one day at a time.' I suggest that the slogan indicates that (former) alcoholics remain aware, all their lives, of the attractiveness of drink (Ainslie, 2000, 80). Moreover, their susceptibility to such an appreciation does not distinguish them from us: we, too, are aware of the attractiveness of options — sexual opportunity, financial impropriety, gastronomic indulgence, or whatever it might be — that we do not judge to be choiceworthy and which we do not choose. The perfectly virtuous agent, who desires only what she judges she ought to, *may* be an ideal agent (I take no stand on the question) but we do not have to aspire to such heights in order to count as autonomous.

Similar considerations block a second reply on behalf of something like a hierarchical view. We might think that the agent does not count as autonomous because her preferences are formed under rationality-distorting conditions. The attractiveness of the drug crowded out other considerations, or cravings and the pain or fear withdrawal clouded the mind. The problem with this line of thought is that it is extremely difficult to specify procedural conditions for preference formation which would rule just the right preferences in and out. Why should we say that addicts fail to properly appreciate the virtues of sobriety, and not that the rest of us fail to appreciate the value of intoxication (after all, the addict has the advantage over of us of having experienced both sides of the question). We might say that the sober life just is better than the life of addiction (and we would be right). But this is a substantive consid-

eration, not a procedural one. If autonomy is self-government, then it is not our values, or even the correct values, that matter: it is the values of the agent herself. When she judges that it is better to consume than to abstain, she fails properly to appreciate the value that sobriety has *for her*, but equally when she decides to abstain she might not fully appreciate the value that consumption genuinely possesses, again for her. At the time she decides to take her drug, she genuinely judges that consumption, either on just this one occasion, or whenever the question arises, is the best thing to do, all things considered; she fails properly to appreciate certain values that are more weighty for her on other occasions, but she also has a livelier appreciation of rival values than on those other occasions.

The problem is not confined to hierarchical theories like Frankfurt's. I suggest that any procedural *synchronic* account of autonomy will be unable to give the right result in cases in which addicts experience judgement shift.<sup>3</sup> Addicts sometimes or often genuinely judge that consumption (on this occasion or regularly) is better than abstention; they may value consumption when the time comes for it, and yet they may still count as autonomy-impaired. They genuinely act against their own will, despite genuinely choosing consumption over abstention. Any account of autonomy which looks to synchronic conflict will not be able to account for the autonomy-impairment in such cases; either the requisite conflict will be missing, or it will be a feature of autonomous as much as autonomy-impaired actions.

The failure of synchronic accounts suggests, naturally enough, that we should turn to diachronic accounts for a solution to our problem. Historical accounts of autonomy are a kind of diachronic approach; do they fare any better here? Recall that on Christman's historical account, a preference is autonomous, *inter alia*, if the agent approves (or would approve) of the manner in which she came to acquire it. This account seems to give us the right result in many cases of unwilling addiction. Addicts may judge that it is better to consume now, given that they experience intense cravings or fear withdrawal, but also judge that it would be better, all things considered, if neither of these were the case: that is, if they were not addicted. But another kind of case is certainly possible, and may actually be very common: under the influence of the drug's attractiveness, the addict may judge that her addiction, which gives her such a lively appreciation of the virtues of the drug (the way it opens the doors of perception, or allows her to take time out in an

---

3 Synchronic theories of autonomy divide into desire-based theories, like Frankfurt's, and judgment-based theories, such as the account advanced by Gary Watson (1975).

increasingly stressful world), is itself valuable or at least no worse than neutral, all things considered. Such an addict will therefore endorse not only her occurrent desire, but also the means whereby she acquired it. Such an agent can nevertheless count as autonomy-impaired, if at other times she sincerely wishes to be free of her addiction.

We need a diachronic account of autonomy, but the existing historical accounts are not adequate for our purposes. They do not give the right result in all cases of unwilling addiction. We need to explain how it can be true that the addict acts against her will, even though she chooses consumption, and values it when she chooses it.

#### IV The Extended Will

I suggest that we identify the agent's will with her *extended agency*. Being an agent, that is, having a single, relatively unified, self, is not something to which we are simply born. Instead, it is an achievement. We gradually unify ourselves. More than a century of psychology, from Freud to cognitive science, has given us good reason to believe that human beings are relatively fragmented. Our minds are built up out of a large number of sub-personal mechanisms, which differ in the extent to which they receive input from each other and from consciousness (assuming there is something to consciousness over and above some subset of the population of mechanisms). Some are 'informationally encapsulated,' which is to say that they are cut off from information from other mechanisms or modules. Subpersonal mechanisms are not merely information processors: they also drive behavior. That is, they not only output to consciousness, but also sometimes bypass it altogether. Thus, conscious perception of what we are doing and why can diverge, under some circumstances, from the real causes of our behavior (Wegner, 2002).

On many views of the mind, the fact that agents seem more or less unified, most of the time, is the real fact requiring explanation. Given that we are constructed out of a disparate collection of relatively autonomous mechanisms, why do we appear, not only to others, but also to ourselves, as a single thing persisting in time? Why is there a self at all? I suggest that unified selves are a result, at least in important part, of negotiation, bargaining and strong-arm tactics employed by subpersonal mechanisms as they attempt to achieve their aims.<sup>4</sup> Each mecha-

---

4 Ainslie stresses the role of bargaining in unifying the self, claiming that it may be 'all that unifies a person' (2001, 43). But, as an anonymous referee points out, the less unified the agent the less opportunity there seems to be for intrapersonal bargaining: the sub-agent with which one bargains may not be around to be

nism requires the cooperation of at least some others if it is perform the task for which it is designed. At very least, it requires that they refrain from interfering with its plans. It therefore needs to bargain with, or constrain, other mechanisms, so that their machinations will not undermine the goals it seeks. As mechanisms engage in this process of bargaining and constraint, a unified self comes into existence. The multiplicity of mechanisms, each pulling in its own direction, comes gradually to be replaced by a self, with a more or less consistent set of preferences, dispositions and desires — in short, a character. This is a process that is never completed, but the major outlines of the self are laid down fairly early in development.

The famous Stanford Marshmallow tests give us some insight into the process by which agents unify themselves. Walter Mischel and his colleagues gave children a choice between an immediate reward — say, one marshmallow — or a larger reward — say, two marshmallows — if they could wait for fifteen minutes. In some conditions, the rewards — smaller, larger, or both — were visible to the children; in some they were hidden from sight. The experimenters expected that attention to the larger reward would concentrate minds and increase the ability to delay. Instead, they found that attention to the rewards diminished delay times. Apparently, physical proximity functions, in the same way as temporal proximity, to cause a crossing of discount curves. Nevertheless, there were significant differences between children, even when rewards were visible. Some children were able to delay for much longer than others. These children were observed to use a number of attention-shifting techniques. They sang, played, covered their eyes, even tried to fall asleep — all, apparently, in a more or less successful attempt to prevent themselves from dwelling on the rewards available to them (Mischel, 1981).<sup>5</sup>

---

punished for defections or rewarded for cooperation. Partly for this reason, I think that strong-arm tactics have a larger role to play in self-unification than does Ainslie. However, intrapersonal bargaining has an important role to play nevertheless. The sub-agents persist across time, and have goals that can be satisfied or frustrated. To that extent, the game is iterated; sub-agents can threaten or promise, in the expectation that the mechanisms with which they bargain will persist to collect their rewards or punishments.

5 Mischel and colleagues call these attempts, by the children, to distract themselves, 'spontaneous attention deployment strategies' (Rodriguez, Mischel, and Shoda, 1989). As an anonymous referee points out, however, some of these behaviors are better seen as the *product* of successfully deploying these strategies, rather than as *instances* of them: the child who succeeds in distracting herself sufficiently from the waiting reward might *as a consequence* fall asleep, rather than falling asleep as a way of distracting herself.

I suggest we see the techniques employed by these children as strategies to extend their will across time. Already by the age of three or four, most children realize that they are subject to preference reversals, which lead them to choose objectively smaller rewards over larger, and they have learnt strategies to prevent these reversals. They apply these strategies with the aim of securing larger rewards, but to the extent they succeed, they achieve a much more important good as a by-product: they increase the extent to which they are unified agents. Their reward-seeking strategies increase the degree of cross-temporal intrapersonal cooperation, as Ainslie might put it: they sacrifice shorter-term interests for longer, and thus make themselves capable of pursuing their own conception of the good. Ainslie suggests we see selves as consisting of nothing more than constellations of interests competing for control of behavior; to the extent to which self-unification strategies succeed, they cut some interests out, and swamp them beneath others. The unified agent is then able to act on her own preferences and values, without fearing that her plans will be short-circuited when the opportunity for some more immediate reward presents itself.

On this view, the unified agent is an achievement. Unification is something we achieve in the process of intrapersonal bargaining, cajoling and coercing, through which each sub-agent attempts to secure the goals it seeks. When we begin the process of shaping ourselves, we do not seek coherence or agent-hood; there is no 'I' to seek these things. As the process continues, however, a self comes into being; it is built out of a coalition of sub-personal mechanisms and processes, but it has sufficient unity to pursue goals and present a single face to the world, and to think of itself as a single being. It is now minimally unified. But unification typically does not cease at this point. From this point on, it — *we* — may continue more or less consciously to work on itself; we shape ourselves in the light of our values and ideals. Sub-agential mechanisms build the self that will then continue to shape itself.<sup>6</sup>

---

6 This view of the self as an achievement has close affinities with the view advanced by Dennett (1991). However, whereas Dennett — in at least some of his moods — seems eliminativist about the self, I see no reason why the constructed self should not count as a *real* self. Dennett has used this account of the creation of selves to explain various pathologies of agency, such as (so-called) multiple personality disorder; on his view, one-self-to-a-customer is merely the typical case, not a fact about the kinds of beings we are. Perhaps some extreme cases of autonomy-impairing addiction result, not so much in a single agent who cannot extend her will across time, but in a fragmentation of the agent sufficient to call into doubt her existence as a single being.

To the extent to which we unify ourselves, and thereby enable ourselves to pursue the goods we genuinely value, we also alter our own discount curves for future goods. We all discount future goods hyperbolically. As we get older, however, and we become more unified agents, our effective discounting rates approach the exponential ideal. Addicts are a partial exception: their discount curves remain highly bowed, so that they experience preference reversals more easily than most of the rest of us (Ainslie, 2001, 34). The very fact that addicts are subject to such reversals suggests that they are less unified than non-addicts.<sup>7</sup> They are unable effectively to exert their will across time. It is in this fact that the impairment of their autonomy essentially lies. A basically autonomous agent is self-governing, and a necessary condition of self-government is the ability to extend one's will across time. The agent who is unable to exert control over her future behavior by shaping her desires and her actions lacks the capacity for self-government. Her preferences at time  $t$ , including her preferences as to how she will behave at  $t1$ , have insufficient influence over her actions at  $t1$ .

In a few short words, then, addiction impairs autonomy — when it does — because it fragments the agent, preventing her from extending her will across time.<sup>8</sup> The typical autonomy-impaired addict *has* a will; she is not as fragmented as all that. But she lacks the capacity effectively to guide her own future behavior by her will. As a consequence of cravings, crowding out of alternatives, or the coercive effects of withdrawal, addicts experience preference reversals which are sharper and less controllable than those to which non-addicts are subject. An unwilling addict is unwilling because, though she chooses to consume, her preference is temporary, and does not reflect her will.

---

7 Ainslie (2000, 80) note that the development of dissociated personalities is more common among addicts than non-addicts; further evidence for their relative disunity.

8 Addiction *need* not impair autonomy. Addicts lack autonomy when they suffer regular and uncontrollable preference reversals, such that they find themselves, when in the grip of their addiction, doing things that at other times they would prefer not to do. The addict who *always* (or usually) prefers consumption does not experience such a preference reversal, and therefore is not autonomy-impaired (whatever else we may think of her).

## V Interlude: Bratman

But what justifies us in identifying the addict's will with her preference to abstain, rather than her preference to consume? It might be helpful to approach this question by comparing the account of autonomy offered here to a closely related view, to which it is heavily indebted: the account of temporally extended agency developed by Michael Bratman. Bratman is concerned with a question similar to ours: what features of agency constitute a person's endorsement of a desire? What features have the authority to speak for the agent? Moreover, like the proposal I am sketching here, Bratman's account looks to the relationships that unify the agent to play this role. He endorses a broadly Lockean view of personal identity, and argues that the states and attitudes which have the role of constituting and supporting the connections and continuities which make up personal identity have the requisite authority to speak for the agent. Roughly, and ignoring various complications, Bratman argues that an agent endorses a desire when she has a self-governing policy, with which she is satisfied, in favor of treating that desire as providing a justifying reason in motivationally effective practical reasoning (Bratman, 2000). Such self-governing policies impose a unity on the agent; because they play an important role in making her the person she is, they have the authority to speak for her.

Bratman has performed a valuable service by rescuing plans and policies from the neglect in which they lay prior to his work. No doubt, he is right in holding that they play an important role in unifying human agency. Nevertheless, we cannot understand the autonomy-impairment characteristic of the addict in terms of plans, at least not as Bratman suggests. Plans and policies, implemented as Bratman envisions, *require* an already unified agent to carry them out; they therefore cannot play the right role in unifying an agent as fragmented as the addict.

Suppose the addict formulates a policy, with which she is satisfied of abstaining from her drug at  $t$ . What reason does she have for thinking that at  $t_1$ , when the time for consumption is imminent, she will not simply dump her policy in favor of a new one (with which she will be equally satisfied)? Resolutions, on Bratman's view, are supposed to preclude reconsideration, but they can do so only if the agent has the habit of non-reconsideration, and such habits presuppose unified agency. Addicts are too fragmented for policies to work. Rather than formulating a policy, addicts need to resort to more direct action. They must prevent the self or person-stage that would jettison the policy from taking control of their behavior, which means either ensuring that their discount curves do not cross or, if they cannot achieve this, that they do not act on their temporary preferences.

Addicts might prevent preference reversals using the kinds of methods we saw employed by the children in the marshmallow test; most simply, by avoiding cues associated with drug-taking which are known to trigger cravings (Loewenstein, 2000, 66; 69-70). Attention direction strategies, such as the self-distraction techniques of the children, are probably the single most commonly employed means whereby agents prevent their discount curves from crossing: the agent who prefers to stick to his diet does not read the descriptions of the available desserts, because doing so will tend to raise the value that eating dessert will have for him; in Ainslie's language, it will tend to increase the extent to which short-term reward-seeking processes have control over his behavior. Self-control consists, largely, not in beating down urges — at least not directly, not by pitting the force of reason against desire — but simply in looking the other way.<sup>9</sup> Addicts are too fragmented for normal attention-distraction techniques to have much chance of succeeding; instead they are most successful when they structure their environments so that the cues which remind them of their drugs are entirely absent. Hence the fact that most of the soldiers who returned from Vietnam addicted to heroin had relatively little trouble kicking the habit, whereas those who try to give up while surrounded by the same people with whom they have consumed find it so very difficult (Loewenstein, 2000, 69-70).

If addicts cannot prevent preference reversals, they can use strong-arm tactics on themselves. In one treatment modality, cocaine addicts write letters confessing their most shameful secrets, to be mailed in the event they drop out of the program (Schelling, 1992, 167). Alcoholics consume the drug disulfiram ('antabuse'), which makes them feel sick if they drink. Both methods aim at raising the cost of consumption, and thereby at strengthening mechanisms sensitive to these costs. In any case, merely formulating a policy is far from sufficient for addicts to achieve the kind of minimal unity and, therefore, autonomy, they seek, and for that reason an account of autonomy-impairment which focuses on such policies misses the extent of their impairment. Addicts fail to stick to their plans and policies, but saying that is saying far too little.

---

9 Richard Holton (2003) argues that the evidence, both phenomenological and empirical, tells against this view. When we exert will-power, it feels like a struggle; moreover agents engaged in self-control tasks show standard signs of physiological effort — increased pulse rate, blood pressure, and so on. But the fact that effort is involved in controlling oneself does not demonstrate that we struggle directly against recalcitrant desires: directing our attention itself takes effort. Self-control is not, or not only, exercised by struggling against desires when they are at their strongest; instead it consists very importantly in the effort to keep them weak.

Their autonomy is far weaker than mere planning failure suggests, because their selves are far more fragmented.<sup>10</sup>

Of course, an addict might well need a policy of exerting strong-arm tactics or attention-shifting methods upon herself. The point is not that policies aren't needed for autonomy; the point is that they are far from enough. Agents become minimally autonomous not by making plans and hoping they'll stick to them; they become minimally autonomous by *forcing* themselves to stick to them. Self-government, like political government, requires a monopoly on the coercive forces of the agent.

Bratman's proposal therefore fails to capture the sense in which addicts (and others subject to predictable and regular preference-reversals: kleptomaniacs, for instance) are disunified. Bratmanesque agency is a more elevated form than that the addict seeks to achieve, at least in the first instance. Indeed, I suggest, it ought not to be understood as an account of *basic* autonomy at all. It is, rather, an account of ideal autonomy, or something well on the way to ideal autonomy. We need not have Bratmanesque autonomy to count as morally responsible agents, agents who are relatively unified and able to answer for themselves. Addicts lack Bratmanesque autonomy, but that is the least of their problems: they lack much more. They lack the capacity to unify themselves to a sufficient degree to begin to formulate plans and policies, in the realistic expectation that they will abide by them.

If Bratman's account of autonomy is too demanding to explain what has gone wrong in the case of the addict, his account of desire endorsement is too permissive to capture the extent to which addicts fail to endorse their desire to consume their drug. For Bratman, a desire is identified with by the agent so long as it is endorsed by her self-governing policies. But addicts are fragmented agents, quite capable of possessing (fragmentary) contradictory self-governing policies. With which set of policies do we identify the agent? On my proposal, the answer is far cruder than on Bratman's: we identify the agent herself with the part-self which is the product of her own strong-arm tactics. Though addicts are fragmentary compared to healthier adults, they are nevertheless relatively unified compared to, say, infants (who are unable to delay gratification at all). An addict has already succeeded in extending her agency

---

10 Similar remarks apply to Holton's account of (what he calls) weakness of the will (Holton, 2003). Holton argues that will power is a separate faculty; he might therefore interpret the autonomy-impairment characteristic of addiction as the result of a pathologically weak will. However, this would be to miss the full extent of the fragmentation characteristic of the addict. She does not fail in the struggle against a recalcitrant desire; instead, she oscillates between endorsed desires.

through time to some significant extent; if she had not, she would not be able to coordinate her actions sufficiently well to acquire and administer her drug. That, forward-planning, agent, is the real self; it is when *that* agent prefers abstention to consumption that she can truly claim that she consumes against her will. By extending our will across time, we make ourselves. Self-governing agents are autonomous agents; to the extent we lack the capacity to govern ourselves, we are autonomy-impaired.

## VI Substantive Conditions of Procedural Autonomy

Clearly, the proposed account of autonomy is procedural. An agent is self-governing just in case she shapes her dispositions and her actions as she wishes. She may shape them well or ill; in the service of noble aims or ignoble. Autonomy is not freedom or self-realization. Autonomy is self-rule, and one can rule oneself well or badly, in desirable circumstances or undesirable. Nevertheless, though it is procedural the view implies some weak substantive conditions.

Consider some examples from the literature of individuals who lack autonomy due to their socialization. First, consider Benson's example of the woman who has been socialized into believing that her self-worth is a function of her attractiveness (Benson, 1991). This *might* undermine her autonomy, but it might not. Autonomy is self-rule; an agent is autonomous to the extent she is able to put her values into effect. So long as she genuinely values physical attractiveness, the machinery of extended agency can go to work, moulding her into the kind of agent she wants to be, based on this value. She makes this value her own, and makes herself, by binding herself to it. I suggest our capacity for taking responsibility for our values through our extended agency goes some way to explaining the common intuition that even agents who have experienced the very worst kinds of socialization — in deeply racist societies for instance — nevertheless cannot use this socialization as an excuse for their failures.

But this view does not require us to reject the genuine insights of proponents of substantive views of autonomy. In fact, it gives us all the resources we need to explain how features of an agent's society can impair the autonomy of unfortunate individuals. Consider Benson's retelling of the *Gaslight* story. In his updated version, sexist science leads a doctor to diagnosis his wife with hysteria, purely on the basis of her active imagination and tendency to emotional outbursts. The protagonist has the misfortune to accept the science that condemns her, As a result, she loses the 'sense of her own status as a worthy agent.' (Benson, 1994, 657). Benson thinks this kind of case shows that accounts of autonomy must be substantive. Perhaps he is right, but if this is so (and

it is basic autonomy that he has in mind), it is only to the extent to which substantive elements fall out of the extended agency account. If we are autonomous to the extent to which we are capable of governing our own actions over time, then anything which undermines the sense that we are capable of governing ourselves is, to that extent, incompatible with autonomy. Sexism, racism and slavery do indeed profoundly impair autonomy; they do so because, as Benson observes, they destroy agents' 'sense of their competence to make their own decisions and manage their own lives' (Benson, 1994, 659). Without such competence, we cannot become effectively self-governing; thus social conditions and oppressive socialization can undermine autonomy understood procedurally.

It is worth remarking that, on this view, autonomy can be impaired by less dramatic social misfortunes than being born into a deeply sexist or racist society. Autonomy is an achievement, and it requires skills of self-control. Lucky agents acquire these skills in early childhood, and are able to demonstrate a fair degree of self-control by around the age of four. Unlucky children are not taught these skills, lack the capacity to develop them, or, perhaps most frequently, find themselves in environments which do not reward delayed gratification (there is no point in holding off on eating that cookie now if one's father will simply take it away and eat it himself [Strayhorn, 2002]). Autonomy can be impaired by a variety of social conditions; oppression, poverty, even inconsistent parenting. Longitudinal studies confirm that agents who fail to develop the skills of self-control tend to fare worse on a variety of measures of success: failure to achieve a high degree of autonomy translates into failure to achieve goals which require delay of reward (Shoda, Mischel, and Peake, 1990). To the extent to which we fail to become autonomous beings, it will be more difficult to engage in any long-term projects, since we shall continually be undermining our own efforts. Autonomy as self-government is merely procedural, and in many ways quite undemanding; but is itself a precondition of pursuing a worthwhile life plan in which substantively valuable goods can be realized.

Philosophers who write about autonomy seem to have widely divergent views of what it consists in, how we get and how we might fail to have it. Some or all of these accounts might prove useful in understanding human agency, across its full range. It may be that ideal or excellent agents have harmonious hierarchical structures of desires; it is certainly the case that the best agents have largely true moral beliefs. No matter how important the aspects of agency analysed by these accounts turn out to be, however, they leave an important condition of even basic autonomy unexplored. Before we can pursue the most valuable goods accessible to human beings, we need to unify ourselves. Only relatively unified agents can pursue long-term goals. We unify ourselves by strengthening certain of our desires and weakening others, so that our

discount curves begin to resemble the exponential curves of classical economics. Agents who cannot achieve this degree of unity are not even minimally autonomous, whether or not they fulfil the other conditions of human agency at its best.<sup>11</sup>

*Received March 2005*

*Revised August 2005*

## References

- Ainslie, G. 2000. 'A Research-Based Theory of Addictive Motivation.' *Law and Philosophy* **19** 77-115.
- Ainslie, G. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.
- Arpaly, N. 2004. 'Which Autonomy?' in *Freedom and Determinism*, J.K. Campbell, M. O'Rourke, and D. Shier, eds. Cambridge, MA: The MIT Press.
- Benson, P. 1991. 'Autonomy and Oppressive Socialization.' *Social Theory and Practice* **17** 385-408.
- Benson, P. 1994. 'Free Agency and Self-worth.' *Journal of Philosophy* **91** 650-68.
- Bonnie, R.J. 2001. 'Addiction and Responsibility.' *Social Research* **68** 813-34.
- Bratman, M. 2000. 'Reflection, Planning, and Temporally Extended Agency.' *Philosophical Review* **109** 35-61.
- Charland, L.C. 2002. 'Cynthia's Dilemma: Consenting to Heroin Prescription.' *American Journal of Bioethics* **2** 37-47.
- Christman, J. 1991. 'Autonomy and Personal History.' *Canadian Journal of Philosophy* **21** 1-24.
- Christman, J. 2003. 'Autonomy in Moral and Political Philosophy.' *Stanford Encyclopedia of Philosophy*, E.N. Zalta, ed. URL = <http://plato.stanford.edu/archives/fall2003/entries/autonomy-moral/>
- Dennett, D.C. 1991. *Consciousness Explained*. London: Penguin Books.
- Dworkin, G. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.
- Elliott, C. 2002. 'Who Holds the Leash?' *American Journal of Bioethics* **2** 48.
- Elster, J. 1999. *Strong Feelings: Emotion, Addiction and Human Behavior*. Cambridge, MA: The MIT Press.
- Fingarette, H. 1988. *Heavy Drinking: The Myth of Alcoholism as a Disease*. Berkeley: University of California Press.

---

11 I am grateful for extremely useful comments from two anonymous referees for the *Canadian Journal of Philosophy*.

- Fischer, J.M. and M. Ravizza. 1998. *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.
- Foddy, B. and J. Savulescu. 2006. 'Can Addicted Heroin Users Consent to the Prescription of Their Drug?' *Bioethics* **20** 1-15.
- Frankfurt, H. 1988. 'Freedom of the Will and the Concept of the Person,' in *The Importance of What We Care About*. Cambridge: Cambridge University Press.
- Holton, R. 2003. 'How is Strength of Will Possible?' in *Weakness of Will and Practical Irrationality*, S. Stroud and C. Tappolet, eds. Oxford: Oxford University Press.
- Holton, R. 2004. 'Rational Resolve.' *Philosophical Review* **113** 507-35.
- James, W. 1890. *Principles of Psychology*. New York: Henry Holt and Company.
- Kane, R. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.
- Leshner, A. 1999. 'Science-Based Views of Drug Addiction and Its Treatment.' *Journal of the American Medical Association* **282** 1314-16.
- Loewenstein, G. 2000. 'Willpower: A Decision Theorist's Perspective.' *Law and Philosophy* **19** 51-76.
- Mackenzie, C. and N. Stoljar. 2000. 'Introduction: Autonomy Refigured,' in *Relational Autonomy: Feminist Perspective on Autonomy, Agency, and the Social Self*, C. Mackenzie and N. Stoljar, eds. New York: Oxford University Press.
- Mischel, W. 1981. 'Metacognition and the Rules of Delay,' in *Social Cognitive Development*, J.H. Flavell and L. Ross, eds. Cambridge: Cambridge University Press.
- Neale, J. 2002. *Drug Users in Society*. New York: Palgrave.
- Rodriguez, M.L., W. Mischel, and Y. Shoda. 1989. 'Cognitive Person Variables in the Delay of Gratification of Older Children at Risk.' *Journal of Personality and Social Psychology* **57** (1989) 358-67.
- Schelling, T.C. 1992. 'Self-Command: A New Discipline,' in *Choice over Time*, G. Loewenstein and J. Elster, eds. New York: Russell Sage Foundation.
- Shoda, Y., W. Mischel, and P.K. Peake. 1990. 'Predicting Adolescent Cognitive and Self-Regulatory Competencies from Preschool Delay of Gratification: Identifying Diagnostic Conditions.' *Developmental Psychology* **26** 978-86.
- Slote, M. 1986. 'Understanding Free Will,' in *Moral Responsibility*, J.M. Fischer, ed. Ithaca: Cornell University Press.
- Sobell, L.C., T.P. Ellingstad, and M.B. Sobell. 2000. 'Natural Recovery from Alcohol and Drug Problems: Methodological Review of the Research with Suggestions for Future Directions.' *Addiction* **95** 749-64.
- Strayhorn, J.M. 2002. 'Self-Control: Theory and Research.' *Journal of the American Academy of Child and Adolescent Psychiatry* **41** 7-16.
- Watson, G. 1975. 'Free Agency.' *Journal of Philosophy* **72** 205-20.
- Watson, G. 1999. 'Excusing Addiction.' *Law and Philosophy* **18** 589-619.
- Wegner, D. 2002. *The Illusion Of Conscious Will*. Cambridge, MA: MIT Press.
- Yaffe, G. 2002. 'Recent Work on Addiction and Responsible Agency.' *Philosophy and Public Affairs* **30** 178-221.

