Autonomy and Addiction

Neil Levy

nllevy@unimelb.edu.au

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 to impair autonomy. This impairment is usually thought to be a simple matter, easily explained by almost any even *prima facie* acceptable account of autonomy. In fact, I shall argue, at least in some cases the impairment is far from straightforward, and none of the theories of autonomy currently on the market are able satisfactorily to explain it. 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.

Accounts of Autonomy

Available accounts of autonomy fall into two broad classes: *procedural* and *substantive*.[1] 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 which 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 rapid 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.[2] 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.

The hierarchical account of autonomy is subject to what some see as a devastating counterexample, from cases of manipulation.[3] 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, 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 and by John Christman.[4] 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.

This kind of worry has especially concerned philosophers who focus upon oppressive socialization. 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 count as autonomous. For instance, Paul Benson argues that only agents who regard themselves as competent to answer for their conduct in the light of normative demands can count as autonomous.[5]

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 use 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.[6] 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 significances 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.

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. This definition may seem to beg the question against substantive conceptions of autonomy. To some extent, the perception that it does so is fair. It is reasonable to think that it is impossible to elaborate a merely procedural conception of human *flourishing*, but it does not follow that an analysis of *autonomy* cannot be content-neutral. If autonomy is just 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. Self-government just is a procedural notion. To that extent, defining autonomy as self-government stipulates

substantive conceptions away. This is not an unhappy result, if I am right in thinking that substantive views have smuggled into their analyses conditions of flourishing agency that go beyond autonomy.

Nevertheless, I shall argue that the conception of autonomy on offer has something to offer friends of substantive conceptions. Though the account is procedural, nevertheless it implies 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.

Addiction.

The debate between the 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 been more or less ignored. The reason why it has played so small a role in this debate 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 impair autonomy, it is widely felt; instead it destroys agency itself.

On this conception, addicts literally can't help themselves. This is 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 comer 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.[7]

This is a view of addiction which is still widely shared, 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'.[8] 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'.[9] 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'.[10] 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 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.[11] There is plenty of direct evidence, in any case, that addicts exercise some degree of control over their consumption behavior. Consumption is price sensitive, in a manner that would be surprising if addictive desires were compulsive.[12] 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 addicts typically go 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 require for a time.[13] 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 varies from drug to drug; cocaine addiction seems to be almost entirely a matter of craving and not withdrawal). But in no cases 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.[14] 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).[15] 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.

This is not to deny that withdrawal may not have a coercive effect upon addicts, or even that those philosophers who have seen in its effects the only excuse addicts have available for criminal actions are wrong; it is simply to suggest that coercion is not the whole story. There is an impairment of autonomy characteristic of addiction, in addition to the coercive effects of withdrawal.

So why do addicts consume their drugs? The short 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 drugs or the money they need to buy them, or why they are likely to readdict themselves after withdrawal. It is not compulsion, or coercion; in some sense, it is 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.[16] But as all of us who have ever struggled with an addiction – whether to

caffeine, tobacco or heroin – know, that it 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 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.

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 judgments of the addict. Most of the time, she 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 convenient explanation of the kind of oscillation here.[17] Economists and psychologists have generally supposed that we discount future goods exponentially. Exponential discounting explains some kinds of inconsistency, but, arguably, not the kind of characteristic of the addict. Suppose the addict discounts both a drug-free existence and the immediate pleasure of consumption exponentially. In that case, the closer in time her opportunity for consumption, the higher her estimate of its value. But, Ainslie points out, exponential discounting does not explain the time inconsistency characteristic of the addict. In particular, it does not explain how it is that the addict can be sincere when she says that she consumes against her will. 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 values consumption more than abstention, and her discount curve is exponential, then at *any* time prior to consumption she will value consumption over abstention. If, therefore, she claims that she acts against her will, she is lying or self-deceived.

Thus, 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 value future good X more than future good Y at time t, and Y more than 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 value X more than 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 weighing 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. 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 pleasure of the drugs 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 seems to her, at the time of action, to be good reasons.

To be sure, the unwilling addict may 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 is 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.[18] Moreover, their susceptibility to such an appreciation does not distinguish them from us: we, too, are aware of the attractiveness of options – sexual opportunities, 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, the pain or fear of withdrawal or the cravings clouded the mind. The problem 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 consideration, 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 to appreciate the value that sobtriety has *for her*, but equally when she decides to abstain she does not 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 to appreciate certain values that are central to her when she does not form that judgment, but 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 at least some cases of addiction.[19] Addicts sometimes or often genuinely judge that consumption (on this occasion or regularly) is better than abstention, and yet 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 well as autonomy-impaired actions.

The failure of synchronic accounts suggest, 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 do 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 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.

The Extended Will

I suggest that we identify the agent's will with *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 an apparently 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 reasons and causes of our behavior.

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 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 strongarm tactics employed by subpersonal mechanisms as they attempt to achieve their aims. Each mechanism 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 its. 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 Mashmallow 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 reward available to them.[20]

I suggest we see the techniques employed by these children as attempt 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 smaller, and they have learnt strategies to prevent these reversals. They apply these strategies with the aim of acquiring 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 strategy increases the degree of cross-temporal intrapersonal cooperation, as Ainslie might put it: they sacrifice shorter-interests for longer. 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 pursue a conception of the good, without fearing that her plans will be short-circuited when the opportunity for some more immediate reward presents itself.

We all discount future goods hyperbolically. As we get older, however, most of us employ various compensatory strategies, so that our effective discounting rates approaches 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.[21] To the extent to which addicts are subject to such reversals, I suggest, they are less unified as agents than are non-addicts.[22] They are unable effectively to exert their will across time. It is in this fact that their impairment of their autonomy essentially lies. An autonomous agent is self-governing; she desires, chooses and acts as she does because she has shaped her dispositions and her behaviours through her choices. The agent who is unable to exert control over her future behavior by shaping her desires and her actions is not self-governed. Her preferences and values at time t do not control her behavior at $t1$ in the right manner.

Interlude: Bratman

It might be helpful to explain this approach to autonomy-impairment by comparing it to a closely related approach to which it is heavily indebted: Bratman's view of temporally extended agency. Bratman is concerned with a question directly related to ours: what features of agency constitute a person's endorsement of a desire? What features have the authority to speak for the agent? And like the proposal I am sketching here, Bratman's account looks to the relationships which 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 constitute personal identity have the requisite authority. 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.[23] Such self-governing policies impose a unity on the agent; because they an important play a 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. 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 comes to formulate a policy, with which she is satisified 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)? 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 the discount curves do not cross or that they do not act on their temporary preferences if they do cross.

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 know to trigger cravings.[24] If they cannot prevent such reversals, they can use strong-arm tactics. 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. Alcoholics consume the drug disulfiram, 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 too say far too little. Their autonomy is far weaker than mere planning failures suggests, because their selves are far more fragmented.

Of course, an addict might need a policy of exerting strongarm tactics or attention-shifting methods upon herself. The point is not that policies aren't needed for autonomy; the point is that they are not enough, in the absence of actual strategic action. 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. By the same token, his account of desire endorsement is too demanding. For Bratman, a desire is identified with the agent so long as it is endorsed by her self-governing policies. But addicts are fragmented agents, quite capable of possessing (fragementary) 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). She has already succeeded in extending her agency through time to some significant extent; if she had not, she would not be able to coordinate her actions sufficiently 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 is unwilling. By extending our will across time, we make ourselves. Self-governing agents are autonomous agents; to the extent we fail at governing ourselves, we are autonomy-impaired.

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, turn to the woman who has been socialized into believing that her self-worth is a function of her attractiveness.[25] 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 our 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 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 society 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 thinks this kind of case shows that accounts of autonomy must be substantive.

Perhaps he is right, but if this is so, 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'.[26]

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 the exercise of skills of self-control. Lucky agents acquire these skills in early childhood, and are able to demonstrate a fair degree of self-control by 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).[27] Autonomy can be impaired by a variety of social conditions; oppression, poverty, even inconsistent parenting. Longitudinal studies confirm that agents who fail to develop these skills tend to 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.[28] 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.

Nevertheless, the substantive conditions of autonomy as self-government remain relatively weak. Agents with systematically false moral beliefs can be autonomous, if their beliefs leave their ability to govern themselves intact. I make no apologies for this limitation of my account; autonomy is a valuable good, but it is only one valuable component of human agency.

NOTES

[1] I borrow this terminology from Catriona Mackenzie and Natalie Stoljar, 'Introduction: Autonomy Refigured', in their *Relational Autonomy: Feminist Perspective on Autonomy, Agency, and the Social Self* (New York: Oxford University Press, 2000), p. 13.
[2] Frankfurt develops his account of autonomy (or, perhaps better, what others have seen as his account of autonomy; whether he intended it as such is a question I leave to one side) in the essays collected in *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988).
[3] Versions of this argument have been advanced by several philosophers, including Michael Slote, 'Understanding Free Will', in John Martin Fischer (ed) *Moral Responsibility* (Ithaca: Cornell University Press, 1986) and John Martin Fischer and Mark Ravizza, *Responsibility and Control: An Essay on Moral Responsibility* (Cambridge: Cambridge University Press, 1998). Frankfurt responds to this charge in his 'Reply to

John Martin Fischer', in Sarah Buss and Lee Overton (eds) *Contours of Agency: Essays on Themes From Harry Frankfurt* (Cambridge, Mass: The MIT Press, 2002).
[4] Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge: Cambridge University Press, 1988); John Christman, 'Autonomy and Personal History' *Canadian Journal of Philosophy* 21, 1 (1991): 1-24.
[5] "Free Agency and Self-worth" *Journal of Philosophy* 91 (1994): 650-668.
[6] See Joel Feinberg, "Autonomy" in John Christman (ed)*The Inner Citadel: Essays on Individual Autonomy* (New York: Oxford University Press, 1989), for an attempt to untangle the many meanings bound up under the single rubric.
[7] James, *Principles of Psychology* (New York: Henry Holt and Company, 1890), p. 543.
[8] Louis C. Charland , 'Cynthia's Dilemma: Consenting to Heroin Prescription' *American Journal of Bioethics* 2.2 (2002),43.
[9] Carl Elliott 'Who Holds the Leash?', *AJOB* (2002), p. 48.
[10] Alan Leshner, 'Science-Based Views of Drug Addiction and Its Treatment', *JAMA*. 1999;282:1314-1316.
[11] See Linda C. Sobell, Timothy P. Ellingstad & Mark B. Sobell, 'Natural recovery from alcohol and drug problems: methodological review of the research with suggestions for future directions'**,** *Addiction* (2000) 95(5), 749–764, for a review of the literature on the recovery of untreated addicts.
[12] Jon Elster, *Strong Feelings: Emotion, Addiction and Human Behavior* (Cambridge, Mass.: The MIT Press, 1999); Joanne Neale. 2002. *Drug Users in Society* (New York: Palgrave, 2002). The same phenomenon occurs among alcoholics: see Herbert Fingarette, *Heavy Drinking: The Myth of Alcoholism as a Disease* (Berkeley: University of California Press, 1988), pp. 36-42. This is the case both within and outside the laboratory; moreover, alcoholics are sensitive to cost even after consuming an initial priming drink.
[13] George Ainslie, 'A Research-Based Theory of Addictive Motivation' *Law and Philosophy* 19 (2000), p. 82.
[14] See, for example, Gary Watson, 'Excusing Addiction', *Law and Philosophy* 18 (1999), 589-619. Watson distinguishes between duress and coercion, and holds that the former, not the latter, is characteristic of addiction.
[15] Richard J. Bonnie, 'Addiction and Responsibility', *Social Research* 68: 3 (Fall 2001), pp. 813-834.
[16] The example comes from two addiction sceptics, Bennett Foddy and Julian Savalescu, for whom addicts are autonomous agents. See Foddy and Savaluscu, 'Can Addicted People Consent To The Prescription Of Their Drug?', *Bioethics*, forthcoming.
[17] See, especially, George Ainslie, *Breakdown of Will* (Cambridge: Cambridge University Press, 2001).
[18] "Even addicts who have been sober for years often say they miss their addictions", Ainslie, 1999, p. 80.
[19] Synchronic theories of autonomy divided into desire-based theories, like that of Frankfurt, and judgment-based theories, such as the account advanced by Gary Watson in 'Free Agency', *Journal of Philosophy*, April 1975, pp. 205-20.
[20] See, for instance, Walter Mischel, 'Metacognition and the rules of delay', in John H. Flavell and Lee Ross (eds), *Social Cognitive Development* (Cambridge: Cambridge University Press, 1981), 240-271.

[21] Ainslie, *Breakdown of Will*, p. 34.
[22] Ainslie note that the development of dissociated personalities is more common among addicts than among others; 'A Research-Based Theory', p. 80.
[23] See, especially, 'Reflection, Planning, and Temporally Extended Agency', *Philosophical Review* 109, 1 (2000): 35-61
[24] George Loewenstein, 'Willpower: A Decision Theorist's Perspective', *Law and Philosophy* 19 (2000), 66; 69-70.
[25] This kind of example is used by Paul Benson in developing a substantive conception of autonomy, in 'Autonomy and Oppressive Socialization', *Social Theory and Practice* 17 (1991), 385-408.
[26] Benson, 'Free Agency and Self-Worth', p. 659.
[27] See Joseph M. Strayhorn, 'Self-Control: Theory and Research', *Journal of the American Academy of Child and Adolescent Psychiatry* 41, January 2002.
[28] There are many studies on the relationship between the extent to which children can delay gratification and later performance on a variety of measures, including SAT scores. See, for instance, Yuichi Shoda, Walter Mischel and Philip K. Peake, 'Predicting Adolescent Cognitive and Self-Regulatory Competencies from Preschool Delay of Gratification: Identifying Diagnostic Conditions' *Developmental Psychology* 26 (1990), 978-986.