Forthcoming in *Noûs*

THE WOODY ALLEN PUZZLE: HOW 'AUTHENTIC ALIENATION' COMPLICATES AUTONOMY[1]

Suzy Killmister
University of Connecticut

ABSTRACT: Theories of autonomy commonly make reference to some form of endorsement: an action is autonomous insofar as the agent has a second-order desire towards the motivating desire, or takes it to be a reason for action, or is not alienated from it. In this paper I argue that all such theories have difficulty accounting for certain kinds of agents, what I call 'Woody Allen cases'. In order to make sense of such cases, I suggest, it is necessary to disambiguating two distinct forms of endorsement, both of which contribute to autonomy.

1.

Anyone who has ever deliberately eaten a twinkie should be able to recognise the phenomenon I am about to describe. You *know* it's disgusting. You know that whatever standards of taste you have, this object fails to measure up. And yet for some inexplicable reason you feel no visceral aversion to the twinkie: instead, you desire it, and you eat it. This experience is not uncommon. We all presumably have fears of things we know to be unthreatening (I, for one, have an inexplicable phobia of karaoke); or feel disgust towards things we know to be benign (many people apparently feel a distinct aversion to the word 'moist'); and sometimes we act on these 'unreasonable' motivational attitudes. We go out of our way to avoid karaoke halls, or typing the word 'moist'. The question I seek to answer in this paper is to what extent such attitudes, and the actions that spring from them, are autonomous.[2]

One way to think about this question is on the model of Frankfurt's unwilling addict: just like the unwilling addict, we might say, we are missing the relevant second-order desire that renders actions free. Since I don't want my fear of karaoke to direct my action, my avoidance of karaoke halls is unfree. This is where the character of Woody Allen comes in.[3] Agents like Woody Allen pose a puzzle for theories of autonomy. The puzzle is this: those motivational attitudes that seem so unreasonable – the desire for a twinkie, the fear of karaoke, disgust at the word 'moist' – can become an integral part of who we are. I can come to define myself, at least in part, as a twinkie-eater, or as a karaoke-phobe.

Even more problematically, not only can the attitudes *themselves* become an integral part of who we are but so too can our *alienation* from those attitudes. In other words, we can come to understand ourselves in terms of, and defend our actions in light of, the unreasonableness of these attitudes. If I see my character as being inextricably neurotic, such that I see desires that arise from that neurosis as expressive of my true self, then not only might I not see the desires' neurotic origin as providing a reason not to act on them, I might in fact see their neurotic origin as providing a positive reason *to* act on them. I'll be using the character of Woody Allen to encapsulate this phenomenon of finding authenticity *through* alienation – of being 'authentically alienated' – before extending the puzzle to different characters.

One of the most striking features of the Woody Allen we see presented in his films, is the attitude he seems to take to his own neuroses. The classic Woody Allen character is one who is driven by desires and anxieties that he knows to have no justification, and yet who takes these desires and anxieties to be reasons to live his life in a certain way. Take, for instance, the character of Isaac in Manhattan. Isaac desires Tracy, a seventeen-year-old high school student. It is at least plausible to read Isaac as recognising that his desire is unreasonable, and yet not only failing to see that as undercutting the justifiability of pursuing a relationship with Tracy, but perversely giving him some *additional* reason to pursue the relationship. Isaac is just that kind of guy – living authentically, for him, just is to pursue these kinds of unreasonable desires.

This puzzle is not merely confined to fictional characters, or trivial decisions such as whether to eat a twinkie. People can – and I would suggest not infrequently do – structure their lives around such 'authentically alienated' desires.[4] For instance, some people are driven by a destructive desire to self-sabotage, whether in the realm of their careers or their intimate relationships (or for some unfortunate souls, both). While this desire can for some people operate at the unconscious level, for others it is a recognised feature of their motivational structure, and one that they can point to not just as a causal explanation for their sabotaging behaviour, but as the reason that (by their lights) justifies their behaving as they do. Such agents self-sabotage in order to self-sabotage – and they can do this without taking their desire for self-sabotage to be reasonable. Or so I will argue.

A key motivation for this paper is to address David Velleman's (2000, 99) demand for a moral psychology that makes room for the "disaffected, refractory, silly, satanic, and punk". We are not all the calculating reasoners so commonly envisaged in action theory, plotting our path through life with one eye on the Good and the other on Reason. My theory tries to make room for less-than-ideal agents such as Woody Allen by neither blithely assuming that all is well and their autonomy is unaffected by their alienation, nor that all is lost and they are straightforwardly non-autonomous.

Instead, I aim to show precisely *which* aspects of their autonomy are affected by such attitudes, and to what degree.

I will argue in Section 2 that many of the standard theories of autonomy do not have the tools to address this puzzle, and are forced to declare either that Woody Allen agents are unproblematically autonomous, or that they are unproblematically non-autonomous. Neither of these pathways is particularly illuminating. Such agents are not straightforwardly non-autonomous, because they are acting in light of what they take themselves to have reason to do. They are not straightforwardly autonomous, because they are acting on the basis of motivational attitudes that are criticisable *by their own lights*. What's needed, then, is a theory that can make sense of this ambiguity. I will offer a sketch of such a theory in Section 3. I will then spend Section 4 unpacking the most relevant aspect of this theory, namely the distinction it draws between endorsing a motivational attitude as reasonable, and endorsing it as a reason for action. It's this distinction that gets to the heart of the Woody Allen puzzle, and thus offers a useful amendment to our thinking about autonomy. In the concluding section I cast the light further afield, considering what the theory says about a more familiar problem for autonomy, namely addiction.

2.

What, then, would some established theories of autonomy have to say about the Woody Allen puzzle? Let's start in the most obvious place, with Harry Frankfurt's (1988) second-order desire theory.[5] According to the second-order desire theory, at least as it was first presented by Frankfurt, an agent is free if the desire that moves her to action is the one that she wants to move her to action. In Frankfurt's terminology, the agent must have a second-order volition that her first-order desire be effective. The classic case of an agent failing this requirement is Frankfurt's unwilling addict: the agent has a desire for the drug, but she does not want this desire to be effective. Nonetheless, she takes the drug. This is to be contrasted with Frankfurt's willing addict, who also experiences the desire, but is satisfied with that desire moving her to take the drug.

Much more could be said about Frankfurt's theory. For the moment, though, this brief sketch is sufficient to illustrate the general problem. Precisely what's puzzling about Woody Allen cases is that such agents commonly *do* want their first-order motivational attitudes to be effective in action. Isaac *wants* to satisfy his neurotic desires; some phobics *want* to avoid their unreasonably fears; some writers *want* to never use the word moist. As such, the basic second-order desire model would have to declare such agents' actions fully autonomous.

The second-order desire model seems to miss something important, in that it fails to capture the psychological ambivalence exhibited by such agents. Michael Bratman (1999) and David Velleman (2000) have each raised a similar concern against Frankfurt's theory, which may prove fruitful for addressing the Woody Allen puzzle. Both note that the kind of second-order volition Frankfurt's theory calls for may be the result of depression or ennui; in such cases, they argue, the agent does not identify with her first-order desires *in the right kind of way* to call her autonomous. Nonetheless, they share Frankfurt's concern with identifying the attitudes or psychological functions that ground agential authority; in other words, the attitudes or functions that speak for the agent.

Both Velleman and Bratman propose similar amendments to the second-order desire model. According to Velleman, "a motive cannot be taken up into the subject's will by just any favourable response to it. It can be taken up into the subject's will only by a favourable response to it as a reason for acting (2000, 13-14)".[6] Bratman similarly construes endorsement of a desire in terms of a self-governing policy "to treat that desire and/or what it is for as a justifying consideration in motivationally effective practical reasoning (2007, 8)."[7] This strategy locates the endorsement necessary for autonomous action in the agent's treatment of the motivating desire as a justifying reason for action.

Unfortunately, the 'reasons-for-action' strategy is no more effective than the second-order desire model in resolving the Woody Allen puzzle. In focusing exclusively on whether such agents take their motivating attitudes to offer justifying reasons for action, these theories fail, just as the second-order desire model does, to capture the ambivalence at the heart of the Woody Allen puzzle. Insofar as Woody Allen agents take their various neuroses to give them reason to act, they will satisfy Bratman's and Velleman's conditions, and thus be declared fully autonomous.

Admittedly, the reasons-for-action strategy can declare a different kind of agent to have reduced. Take Sally, who has a fear of flying. She believes that this fear is irrational – she knows that planes are far safer than cars, which she happily drives every day. Nonetheless, when she is contemplating where to go for her next holiday, she takes her fear of flying to be a decisive consideration and thus chooses to drive. Note that on a simple second-order desire model, Sally's action would be considered autonomous: she has a second-order volition to act on her fear. However, Bratman (2007, p.37) stresses that a desire that motivates "only because the agent aims at getting rid of [i.e. alleviating] the desire" shouldn't count as endorsed in the right way. The agent may endorse the desire in a limited sense as an effective motive, but she does not endorse it in the necessary sense as providing what Bratman calls a 'justifying end'. I take it that what Bratman is suggesting here is that we can draw a distinction between different attitudes we might take towards a desire, even while we endorse it in Frankfurt's sense. On the one hand, we might reconcile ourselves to its existence, and shape our lives

around it accordingly; on the other hand, we might take the desire to be tracking something which we endorse as an end within our schema of values and commitments. It's only the latter form of endorsement that secures autonomy, on Bratman's account.

To determine whether Sally satisfies the reasons-for-action model of autonomy, then, we would need to know more about *why* Sally takes her fear to be decisive. If it is just a matter of avoiding the emotional trauma of invoking her fear, then her situation will be analogous to the addict who simply wants relief from her desperate desire for the drug: she will not be treating the fear as a 'justifying end', and thus will not be acting autonomously.[8] We can, however, redescribe Sally's scenario such that her fear is, in fact, a 'justifying end'. Like the Woody Allen character, Sally may see her fear as an important part of who she is, such that she would, in some sense, be failing to be true to herself if she refused to treat it as an end. For agents who value a certain kind of authenticity, strong emotions can be seen as a kind of personal compass, indicating the direction their life should take. If Sally were one of these kinds of agents, she would not take her fear to be something to be overcome, but rather to be an end that she should have a self-governing policy of promoting. In such a situation, the reasons-for-action strategy would have to consider Sally unproblematically autonomous – the motivating attitude that moves her is one that she takes to be reason-giving. This strategy thus blinds us to the problematic fact that Woody Allen agents such as Sally take their motivating attitudes to be *criticisable*. This introduces an important ambivalence into their actions, an ambivalence that should be captured rather than obscured by a theory of autonomy.

Recall that a primary motivation for Frankfurt, Bratman and Velleman was to identify the attitudes or mechanisms that constitute agential authority. The difficulty is that for Woody Allen cases, the ambivalence does not come from a motivation external to the agent's viewpoint, but rather stems from the fact that the agent's viewpoint itself appears suspect. That is, the agent's practical standpoint seems to be compromised, insofar as it incorporates attitudes the agent has deemed unreasonable.

This may seem to suggest that we should refocus our attention from agential authority to something like alienation: rather than asking whether the desire is reason-giving, we should perhaps ask whether the agent is alienated from the desire. Consider, for instance, the account of autonomy developed by John Christman (1991; 2007; 2009): for Christman, an agent is autonomous insofar as she would not feel alienated from her desire, were she to know of its origin.[9]

What would such a theory say about Woody Allen cases? Well, insofar as Woody Allen agents find their motivational attitudes to be unreasonable, it is perhaps plausible to say that they feel alienated

from them. Insofar as they experience particular motivational attitudes as a force to which they are subject, and which they can't reconcile with other aspects of their self-understanding, then we may consider them to be alienated from those attitudes. If lack of alienation were considered the primary criterion for autonomy, then at least some Woody Allen agents would need to be deemed non-autonomous. However, such a determination risks overlooking the importance of the agential authority that is identified through the reasons-for-action strategy: Woody Allen agents may feel alienated from their motivational attitudes, but they nonetheless take them to be reason-giving. Surely this should factor in our determination of their autonomy.

At this point, it may be objected that I have been uncharitable to the above theories. In arguing that they are unable to capture the nuances of the Woody Allen puzzle, I have been assuming that these theories must characterise agents as either autonomous or non-autonomous. Yet if the purpose of these theories is to develop necessary and sufficient conditions for *paradigmatic* cases of autonomy (c.f. Velleman 2000, 189), then it should be possible to tweak them to allow for cases of less than full autonomy. So if we returned to the second-order desire model, we could perhaps say that agents such as Woody Allen only *partially* endorse their motivating attitudes, and are thus less than fully autonomous. Or we could return to the reasons-for-action strategy, and say that such agents only take their motivating attitudes to be *attenuated* reasons, and thus their autonomy is compromised. Likewise, we could return to the alienation strategy and say that Woody Allen agents are not *fully* alienated from their motivating attitudes.

As tempting as this response may be, it is nonetheless insufficient. The difficulty of the Woody Allen puzzle isn't just that such agents seem to be less than ideally autonomous, it's rather that they seem to be more autonomous along some dimensions, and less autonomous along others. Moreover, both of these dimensions relate directly to the question of endorsement: such characters endorse their motivating attitudes in one sense, but not in another. They are critical of the attitude, but accept its role in their practical reasoning. Insofar as endorsement is understood as a singular phenomenon – as it is in a second-order desire theory, and in a reason-for-action theory, and in a theory of alienation – we will not be able to capture the complexity of the Woody Allen puzzle.[10] The solution to the puzzle, then, involves disambiguating two distinct forms of endorsement, both of which contribute to autonomy.

3. An Alternative Theory

I will now take a slight detour in order to outline a theory of autonomy that I believe has the resources to capture the complexity of the Woody Allen puzzle. The fundamentals of the theory are as follows.[11]

Autonomy, at its most basic, is concerned with the will; that is, with the springs of our action. This concern has two key focal points. First, there is a question about the quality of the will. Do our motivating attitudes reflect who we truly are? Do they issue from us, rather than impose themselves upon us? And second, there is the ability to translate our will into action. To capture this basic idea, I propose a theory according to which autonomy is constituted by three distinct dimensions. To put it slightly differently, if we were to attempt to determine the extent to which an agent was autonomous with respect to a particular action or domain, there would be three distinct dimensions we would need to take into account.

The first dimension is Self-Definition, and it focuses on the quality of the agent's will. This is to be understood in terms of a specific kind of endorsement: an agent is considered Self-Defined on my account insofar as the motivational attitudes she is subject to are ones that she takes to be reasonable (where reasonable is a technical term to be defined shortly).[12] As such, each of the three agents presented in the introduction (i.e. the twinkie-eater, the karaoke-phobe, and the moist-avoider) exhibit low Self-Definition. Each is subject to a motivational attitude – desire, fear, and disgust, respectively – that she takes to be unreasonable.

Being even maximally Self-Defined, however, still leaves our autonomy vulnerable. Since self-governance requires not only governance *of* ourselves, but also governance *by* that self, it requires further that our will find expression in action. I break this requirement down into two distinct dimensions. First is the dimension of Internal Self-Realisation, which measures the extent to which the intentions we form correspond to, and are brought about by, what we take ourselves to have reason to do. Whether or not our Woody Allen cases exhibit high Internal Self-Realisation depends, then, on whether their intentions to eat the twinkie, or avoid the karaoke hall, or avoid typing the word moist, accords with what they take themselves to have most reason to do.

The final dimension is External Self-Realisation, which measures the extent to which the actions performed accord with, and are brought about by, our intentions. So returning to our Woody Allen cases, we would need to check whether, for instance, the agent formed an intention *not* to eat the twinkie, but nonetheless found herself eating it. For the purposes of this paper, only the first two dimensions will be relevant. I will argue that getting clear on the difference between Self-Definition and Internal Self-Realisation, and in particular the distinct roles that endorsement plays in each, is the key to understanding the Woody Allen puzzle.

For Self-Definition, endorsement of an attitude is understood in terms of whether the agent takes that attitude to be reasonable. For instance, an agent may take her fear of mice to be unreasonable, on

the grounds that mice are not conceivably dangerous. By contrast, Internal Self-Realisation casts endorsement in terms of reasons for action. Internal Self-Realisation asks whether the agent's intention tracks what she takes herself to have most reason to do, given her motivational attitudes and beliefs. As such, an agent's intention to avoid mice would only achieve a high level of Internal Self-Realisation if she took herself to have good reason to avoid mice. The key to unlocking the Woody Allen puzzle lies in the claim that an agent can take herself to have a reason to act on the basis of a motivational attitude that she takes to be unreasonable. So an agent can take herself to have a reason to avoid karaoke halls, even though the only motivation to avoid karaoke halls is a fear that she takes to be unreasonable. Likewise, an agent can take herself to have a reason to eat a Twinkie, even though she takes her desire to eat the Twinkie to be unreasonable. To see how this is the case, more needs to be said about precisely what is meant by both Self-Definition and Internal Self-Realisation. I will start by examining the idea of reasonableness being appealed to in Self-Definition.

As noted above, I am using 'reasonable' as a technical term, to pick out a particular kind of endorsement we can bestow upon our motivational attitudes. It is intended to be a relatively weak form of endorsement, in that it is satisfied through the absence of a judgment of unreasonableness: we take an attitude to be reasonable when we don't take it to be irrational, or unfitting, or morally inappropriate, and so on. What needs to be stressed at the outset, in addition, is that the particular way of cashing out unreasonableness (i.e. as irrational, or unfitting, or morally inappropriate) is to be determined subjectively. That is, for any given agent in any given context, it is up to her what narrower standard justifies a judgment of reasonableness. Reasonableness is thus an umbrella term, covering whichever narrower standard the agent applies in a given case.

This can perhaps be better understood by considering some of these narrower standards. One standard via which a judgment of reasonableness can be made is that of 'fittingness', which in turn can be understood in terms of whether the motivational attitude is held for the 'right reasons'.[13] To take a stock example, a fear may be deemed fitting if it emerges in response to something threatening, like a person with a gun, but not fitting if it emerges in response to something harmless, like an infant. Likewise, disgust may be deemed fitting if it emerges in response to something vile, like vomit, but not fitting if it emerges in response to something pleasant, like a chocolate cookie. As such, one of the ways in which an agent may deem an attitude unreasonable is if she takes it to be unfitting. For instance, Sally may deem her fear of flying to be unreasonable on the grounds that flying is not in fact dangerous (or at least, no more dangerous than activities that she is not remotely afraid of).

An alternative standard via which a judgment of unreasonableness could be made is that of moral appropriateness. A motivational attitude may be deemed unreasonable, then, insofar as the agent

takes it to be morally inappropriate to experience it. For instance, an agent might judge that the envy she feels at the success of a close friend is unreasonable, insofar as she believes it is morally inappropriate to feel that way.

The crucial move for understanding Self-Definition is to see that the reasonableness of a motivational attitude is set subjectively, and on two distinct levels. First, the standard the agent invokes in her judgment of reasonableness is set subjectively. Take two agents, both of whom experience anger towards a young child. One may take her anger to be reasonable, because she takes it to be fitting (the child has deliberately done something hurtful, and the agent takes it to fitting to respond to hurtful actions with anger). The other agent, meanwhile, takes her anger to be unreasonable, on the grounds that she takes it to be morally inappropriate to experience anger towards a young child. On my account, the former agent is more Self-Defined than the latter. All that matters for an agent's Self-Definition is whether the standard the agent invokes is satisfied for the relevant attitude; that attitude is not required to satisfy any further standards.[14] Reasonableness is thus a significantly weaker requirement than other concepts in the area, such as rationality, or justifiability, or consistency. Even for a single agent, different attitudes will prompt different standards: while the agent might invoke rationality as a relevant standard for her fears, she may not consider rationality relevant to determinations of the reasonableness of love. Similarly, some agents may take consistency of attitudes to be vital to determinations of reasonableness, while for other agents the inconsistency of her desires will not be sufficient to invoke a judgment of unreasonableness.

The second level on which reasonableness is subjective relates to the criteria of success for the invoked standard. So assuming that our agent invokes the standard of fittingness to justify her fear of mice, it does not matter whether that fear is in fact objectively fitting; all that matters is whether the agent *takes* it to be fitting. This double subjectivity is necessary to avoid the theory of autonomy collapsing into a theory of orthonomy, or 'right' rule (but c.f. Wolf 1993). The standard we should be seeking isn't one that demands the agent be ideally responsive to *good* reasons, but instead that she be responsive to *her* reasons. Unreasonable motivational attitudes are only problematic for *autonomy* insofar as they are criticisable from the agent's own perspective, since it's only in such cases that we can say she has failed to fully govern herself.[15]

Two clarifications are required before we move on to Internal Self-Realisation. First, talk of the agent invoking a standard for her motivational attitudes may suggest the need for a conscious articulation of that standard, and a conscious reflection on whether her attitude meets that standard. However, this would be to build too cognitive a picture of autonomy. I am assuming that for most of us, most of the time, these standards are largely implicit. I do not need to be able to articulate what my

standard of reasonableness for fear is, for instance, to recognise when I am afraid of something that doesn't meet that standard. In other words, I can recognise *that* something is wrong without being able to say what *makes it* wrong. For the most part, I suggest, we simply find ourselves judging our motivational attitudes to be either reasonable or unreasonable, and there is no reason to think that we must be able to stand back from these judgements, and articulate what they are tracking. Nonetheless, the account does require a degree of self-awareness on the part of the agent. As I will go on to argue in Section 4, the reason why unreasonable motivational attitudes reduce autonomy is that they are criticisable from the agent's perspective; their persistence thus indicates a failure to govern oneself in accordance with one's own rules. The agent thus needs to able to recognise *that* she finds an attitude unreasonable, even if she can't articulate precisely *how* it falls short of being reasonable.

Second, the idea of taking a motivational attitude to be reasonable needs to be clearly distinguished from having a second-order desire for that motivational attitude to persist.[16] There are many motivational attitudes that it is unpleasant to experience – fear, disgust, antipathy – and which the agent may thus prefer not to experience, yet which she nonetheless takes to be reasonable. The key question for Self-Definition is not whether the agent believes whether she would be better off without having the attitude, or desires to get rid of it, but rather whether she takes the attitude to fail to satisfy a subjectively relevant standard.

With this account of reasonableness in place, we must now consider what it means for something to be a reason for action. The first distinction that needs to be stressed is between motivating reasons and justifying reasons. My concern is with whether the agent takes the motivational attitude to justify her action, as opposed to seeing it as a causal explanation for her action. A motivational attitude may fall short of justifying action in one of two ways: it can either be outweighed, or it can be bracketed. For a particular motivational attitude *x* to be outweighed, it merely needs to be the case that there are other motivational attitudes that both conflict with *x*, and that are taken to provide sufficient reason not to act on *x*. For instance, I may have a desire to go to the pub, and a conflicting desire to continue working on this paper. Assuming I take my desire to finish the paper to provide a weightier justification than my desire for a beer – perhaps because it meshes more fully with broader plans and intentions – my desire for beer would not justify abandoning my work.

What is important to note about such cases is that the agent retains a belief that the original motivational attitude is reason-giving to some degree, but rejects it as sufficiently weighty to justify a particular action. By contrast, a motivational attitude is bracketed in a particular instance when the agent rejects it as reason-giving altogether.[17] For example, I may experience a sudden desire to eat a

twinkie, despite not being remotely hungry, and despite the knowledge that I do not remotely enjoy the taste of twinkies. I know, moreover, that the only explanation for my sudden desire is that I have just overheard a jingle advertising twinkies. In such a case, I may take my desire for a twinkie to provide me with *no* reason whatsoever to eat one. It is not that this desire is outweighed – perhaps by concerns for my health – but rather that the desire is denied *any* weight in my consideration of what to do.

Just as for reasonableness, reasons for action – at least insofar as they are relevant to autonomy – are subjective, on my view. Whether a particular motivational attitude is outweighed or bracketed will depend upon the agent's preference orderings and commitments. As such, this discussion of reasons for action is not intended to bear any relation to the internalism/externalism debate about reasons more generally. My argument is orthogonal to that debate, in that it is silent on the question of what reasons the agent has, all things considered, and concerns itself exclusively with what reasons the agent *takes herself* to have.[18] What Internal Self-Realisation measures, then, is the extent to which the agent's intentions map onto the reasons she takes herself to have. Insofar as she forms an intention to fulfil a desire that she takes to be outweighed or bracketed, she fails to act fully autonomously.[19]

The key claim, then, is that there is a conceptual gap between taking a motivational attitude to be reasonable, and taking it to be a reason for action. It is this gap that the reasons-for-action model as developed by both Bratman and Velleman fails to heed. Sally takes her fear of flying to give her a reason for action, while at the very same time taking that fear to be unreasonable (because unfitting); the same applies to the twinkie-eater. While the gap can operate in both directions – an agent could take a motivational attitude to be reasonable, and reject it as a reason for action; or she could reject the attitude as reasonable, while taking it to be a reason for action – it is the latter of these that is at stake in the Woody Allen case.

Even if this claim is granted, however, it still leaves open an important question: while conceptually distinct, is it psychologically plausible for an agent to take a motivational attitude to be a reason for action if she also takes it to be unreasonable? I think the answer to this is an emphatic yes.[20] What Woody Allen cases illustrate is that, for at least some agents, reasonableness does not exhaust what's reason-giving. Woody Allen agents recognise that their relevant motivational attitude is unreasonable; nonetheless, they find an alternative pathway to justify – at least to themselves – that the attitude is reason-giving. This is the idea of 'authentic alienation': while Woody Allen agents take some of their motivational attitudes to be unreasonable, those very same attitudes form a core part of how they understand themselves. These attitudes are 'authentic', in the sense that the agents take

them to express a deep truth about themselves. Acting on those attitudes, then, is a form of self-expression, or self-realisation. This is where the reason-giving force of those attitudes is generated.

Woody Allen cases, and the 'authentic alienation' that they embody, provide an example of one particular motivational pathway via which an agent may come to have what looks at first glance like a very unstable psychological profile. However, Woody Allen agents are not the only kinds of agents for whom reasonableness and reasons-for-action can come apart. One alternative pathway might be characterised as ironic detachment. While the Woody Allen agent identifies deeply with the unreasonable motivational attitude, and it is on this basis that she takes the motivational attitude to be reason-giving, the ironically detached agent is less invested in the relevant attitude. An example of such a character would be a modification of the twinkie-eater with which we started. Such an agent needn't take twinkie-eating to define who she is; instead, she might view her desire for a twinkie with a kind of detached curiosity. Precisely because it feels alien, and so out of character, she might take herself to have a reason to indulge its fulfilment.

A third possible pathway could be seen as a form of masochism. Rather than taking the rogue motivational attitude to be reason-giving on the grounds that it is authentically her own, as it is in the Woody Allen case, the masochist shares with the ironically detached agent the view that the attitude is reason-giving precisely because of its lack of fit with who she takes herself to be. Unlike the ironically detached agent, however, the reason-giving force of the desire for the masochist is a form of punishment. The masochist agent punishes herself for her criticisable motivational attitude by forcing herself to bear the brunt of it. Plato's Leontius could be an example of such a character. As Plato (1974, 439e) describes the case, Leontius responds to his desire to look at recently deceased corpses by "[running] up to the corpses, opening his eyes wide and saying to them, 'There you are, curse you – a lovely sight! Have a real good look!'". So while the Woody Allen agent may say to herself: 'I know that my desire is unreasonable; but the desire is mine so I should act on it', the masochistic agent says instead: 'I know that my desire is unreasonable; I should therefore inflict its full implications on myself.'

To summarise the argument so far: I have claimed that current theories of autonomy are unable to fully engage with the problems raised by Woody Allen agents, because they construe the endorsement necessary for autonomy in one-dimensional ways, either as lack of alienation from a desire, or as taking that desire to be reason-giving. The fruitfulness of the theory of autonomy presented here lies in the space it opens up between an agent's judgment of the reasonableness of her motivational attitude, and her having a reason *to act on* that motivational attitude.[21]

4. Defending Self-Definition

My approach to the Woody Allen case relies on it mattering for the agent's autonomy that she is subject to motivational attitudes that she takes to be unreasonable. Since this is not necessarily an obvious claim, I will consider here three possible avenues by which it might be challenged, and argue that none succeed in undermining the core argument

First, it might be objected that unreasonable motivational attitudes are subject to a self-correcting mechanism. For instance, once we realise that the object of our fear is not the right kind of thing to be afraid of, we will cease to fear it. Scanlon, for instance, claims the following: "when a rational creature judges that the reasons she is aware of count decisively against a certain attitude, she generally does not have that attitude, or ceases to have it if she did so before (1998, 24, c.f. 20)." D'Arms and Jacobson also come very close to something like this claim:

> Although it is possible to be afraid of something without judging it fearsome, or to judge it fearsome without actually fearing it, these are uncommon and unstable combinations. […] Such conditions put psychological and rational – that is, causal and normative – pressure on us to alter our feelings or our judgments in order to bring them into harmony (2000, 67).

If this were the case, we might think, reduced Self-Definition would be at most a fleeting phenomenon, and thus not worthy of inclusion in an account of autonomy.

The key problem with this objection is its psychological implausibility. Motivational attitudes are often stubbornly resistant to what we might think of as counter-evidence. An example should hopefully suffice to illustrate the point. A recent science exhibit called 'Sip of Conflict' is set up to explore just this tension. The exhibit invites passers-by to drink from a water fountain. The twist is that the fountain is mounted in a (clean, unused) toilet bowl. Video footage of people engaging with the exhibit clearly demonstrates the kind of prolonged tension that D'Arms and Jacobson's account claim is uncommon and unstable: they experience disgust, and that disgust persists despite explicitly acknowledging that its object is not, in fact, disgusting.[22]

In light of this response, a second objection might be raised, namely that Self-Definition is too stringent to be appropriate as a component of autonomy. Most agents will frequently experience motivational attitudes that don't meet their general standards of reasonableness – indeed, some agents may seek out situations that have just such a predictable effect. Consider an agent who enjoys watching zombie movies, precisely because of the visceral thrill of experiencing fear in response to something that is not in fact threatening. Surely we shouldn't be committed to saying that such people are only

autonomous insofar as they have a standard of fittingness that is loose enough to include fictional creatures as appropriate objects of fear.

The problem with this purported counterexample is that it misrepresents the scope of reasonableness. The idea is not that every agent is taken to have a fixed standard of reasonableness (cashed out in some combination of fittingness, moral appropriateness, etc), and then all of her motivational attitudes must be checked against that standard. Rather, the agent's standards of reasonableness can be flexible and context-dependent. What she takes it to be reasonable to feel in a movie theatre may be very different from what she takes it to be reasonable to feel lying in bed at night. So our agent may take her fear of the zombie to be perfectly reasonable during the movie, and this would be compatible with her taking it be unreasonable if it persisted once the movie was over, and she emerged into the cold light of day.

It is of course possible that there is some agent for whom feeling fear during a zombie movie *is* autonomy-reducing, i.e. if their standard of fittingness is indexed to objective threats. A parallel case may make this clearer. I have (what I take to be) an uncharacteristic tendency to weep at cloying romantic comedies.[23] Insofar as I find this troubling, it is not because of any general objection to feeling emotions in response to fictional characters. Rather, I am annoyed with myself because I have some kind of standard – admittedly largely unarticulated – regarding which kinds of situations it is appropriate to respond to with sadness, and which it is not. To find myself feeling sad at a situation that I know to be both deeply implausible and mind-achingly trivial invokes a kind of cognitive dissonance. I will say more about the importance of this dissonance in response to the final objection.

Rather than claiming that Self-Definition is irrelevant because self-correcting, or overly-demanding because so common, our final critic might object that it is simply irrelevant. Autonomy is about doing what we take ourselves to have reason to do. Why think that our motivational attitudes need to meet any kind of further standard? This objection is particularly salient given one of the implicit commitments of the theory, namely that autonomy can be reduced merely through the experience of a motivational attitude, even if that motivational attitude plays no further role in the subject's practical reasoning.

More must be said, then, to explain why Self-Definition is an important component of autonomy. Self-Definition is a necessary component of autonomy, I take it, because our attempts at self-governance do not merely point outwards, through our attempts to realise our will in the world, but also point inwards through our attempts to shape who we are. Autonomy is not merely governance *by* the self, but also governance *of* the self. This does not call for the impossible task of self-creation, but it does

require us to have a modest moderating capacity. Insofar as we are subject to motivational attitudes that we take to be unreasonable, we exhibit a psychological profile that is criticisable by our own lights. Insofar as we fail to rid ourselves of such attitudes, then, we are failing to uphold our own standards – we are failing to be fully self-governing.

When an agent takes a motivational attitude to be unreasonable, she is making an implicit judgment that this aspect of herself falls short of some standard. Moreover, the standard that the attitude falls short of is her *own* standard, and her judgment of unreasonableness is the upshot of taking that standard to be applicable in the given situation. Even if that motivational attitude never receives uptake in her deliberations about what she has reason to do, then, the persistence of that attitude exhibits a failure to bring aspects of her self into line with her own implicit judgment of how she should be. As such, unreasonable motivational attitudes evince a failure of self-governance. This is why Woody Allen agents are not unproblematically autonomous.

Admittedly, given that in our examples these agents are portrayed as acting on the basis of their unreasonable desires, it is better for their autonomy that these attitudes have been taken up as a reason for action. Otherwise, they would be acting in opposition to what they take themselves to have reason to do, which is quintessentially akratic action.[24] Nonetheless, this 'taking up' of the attitude as a reason for action does not erase the fact that they are acting on the basis of a motivation that they are committed to rejecting, and this means that they are less autonomous than they otherwise could be, if they were acting on the basis of attitudes that they deemed *both* reasonable and reasons for action.

5. Conclusion

I have tried to show that the theory sketched here has the resources to better understand autonomy, in light of its ability to make sense of Woody Allen agents. I hope that this is sufficient reason to consider this theory of autonomy a viable contender. Nonetheless, in concluding, I will attempt to bolster this claim by considering whether the theory can also shed light on a more common problem for autonomy, namely addiction.

Frankfurt's early work was instrumental in prompting consideration of the relationship between addiction and autonomy.[25] Crucially, Frankfurt claimed that addiction was not necessarily in conflict with freedom of the will – what matters is not whether the agent is addicted, but rather whether she endorses her addiction. So we can distinguish between, on the one hand, willing addicts, who want their desire for the drug to move them to action; and, on the other hand, unwilling addicts, who

experience the same (more or less) irresistible first-order desire for the drug, but who do not want that desire to move them to action.

What, then, does the theory presented here say about such cases? Translated into the terms of the theory, we can describe the fully autonomous addict as follows: she accepts her desire for the drug as reasonable (high Self-Definition); she takes her fulfilment of that desire to be in line with what she has most reason to do, and thus forms an intention to consume the drug (high Internal Self-Realisation); she then acts on the basis of the intention and consumes the drug (high External Self-Realisation).[26] Where the theory proves illuminating is in considering what to say about the unwilling addict.[27] While for Frankfurt this determination turned simply on whether the agent wanted her desire to be effective in action, the three-dimensional theory allows us to ask a more nuanced question about the agent's relationship to her desire, thus allowing us to make a further division within the category of unwilling addicts.

To determine what the three-dimensional theory would say about an unwilling addict, we need to know not just whether she takes her desire for the drug to be a reason for action, but also whether she takes her desire for the drug to be reasonable. As the Woody Allen puzzle has shown us, these two determinations can come apart. For some unwilling addicts, then, the unwillingness enters only at the level of reasons for action. That is, such akratic addicts accept that their desires are reasonable (i.e. that the drug is, by their own lights, desirable), but nonetheless reject that desire as a reason for action. For instance, an addict may reject the desire as justifying drug-taking on the grounds that it is vastly outweighed by competing reasons, such as protecting her health, employment, intimate relationships. To nonetheless form the intention to consume the drug is to exhibit compromised Internal Self-Realisation: the intention is not tracking what the agent takes herself to have most reason to do.[28] Compare this to another agent, who not only rejects her desire for the drug as a reason for action, but furthermore takes the desire itself to be unreasonable.[29] Such an agent is caught in the even more troublesome trap of acting on the basis of a desire that has been rejected on two distinct levels. We might call such an agent a 'doubly unwilling addict'. Standard theories of autonomy as endorsement cannot differentiate between the unwilling and the doubly unwilling agent, and they thus overlook the additional reduction in autonomy of the latter agent.[30]

For both addiction and Woody Allen cases, the theory presented here captures a distinction that is frequently overlooked. Once we see that it matters *both* whether the agent endorses her motivational attitudes as reasonable, *and* whether she endorses her motivational attitudes as a reason for action, we open up important space for a nuanced analysis of precisely *how* an agent's autonomy may be reduced when she fails to live up to the ideals of moral psychology.

[1] Many thanks to Sonya Charles, Anita Superson, Paul Bloomfield, David Ripley, and especially two anonymous reviewers, for their generous and helpful comments on earlier versions of this paper. Thanks also to the audience at the Australasian Association of Philosophy Conference in Brisbane 2013, where the ideas in this paper were first developed.

[2] It may seem at first glance that these examples are importantly different: there is something inexplicable about eating something we deem disgusting, while avoiding that which we fear (even if that fear is ungrounded) often seems eminently reasonable. Nonetheless, as will hopefully become clear in Section 3, these examples share an important structural feature: in each case, the agent experiences a motivational attitude that she takes to be unreasonable, which we might think undercuts its reason-giving force. Nonetheless, in acting on it, she treats it as reason-giving.

[3] I should stress that my focus on 'the character' of Woody Allen is meant to be distinguished from a focus one might have on Woody Allen the person. My interest here is solely in the presentation that Woody Allen offers of himself through his films, however much that may diverge from his actual persona.

[4] As an anonymous reviewer for this journal pointed out, the idea of the 'tortured artist' also fits this profile.

[5] Frankfurt originally developed this theory to account for freedom of the will, rather than autonomy. Nonetheless, his theory has been the springboard for many contemporary theories of autonomy, and so deserves mention here.

[6] It is worth noting that Velleman's usage of the term 'will' differs from Frankfurt's here. For Frankfurt, the will refers to an effective desire, i.e. one that moves the agent to action (1998, 14). For Velleman, by contrast, the will is the 'faculty of intentions' (2007, 210), and is guided by reasons. While for Frankfurt acting in accordance with our will may actually undermine our autonomy (where we do not have the will that we want), I take it that for Velleman acting in accordance with our will is, *ceteris paribus*, autonomy enhancing.

[7] For an outline of the differences between Bratman and Velleman's views, see (Bratman 1999, pp.203; Velleman 2007).

[8] Relatedly, agent's such as Sally may become self-deceived, in that they initially experience their fear as unreasonable, but respond by convincing themselves that it is in fact justified: planes fall out of the sky; I have more control over a car, etc. (Thanks to Anita Superson for raising this possibility.) I take this kind of self-deception to be problematic for autonomy, in ways that aren't necessarily captured by the theory that I will go on to sketch. I say more about the problem of deception in n.16.

[9] It should be stressed that Christman's theory of autonomy is more complex and nuanced than I can present it here. Importantly, for Christman alienation involves an inability to reconcile the desire with the agent's narrative understanding of herself, rather than a simple dissatisfaction.

[10] I have presented the problem here in terms of 'reasons-for-action' theories, on the one hand, and alienation theories, on the other. These are not, however, the only classes of theories that are available; more importantly, nor are they the only classes of theories that are susceptible to this challenge, suitably reframed. Take, for instance, the range of theories that construe autonomy in terms of control conditions (see, i.e., Mele 1995a; Fischer and Ravizza 1998). For Mele, an agent is self-controlled (a necessary condition for autonomy) insofar as her intentional action coheres with her better judgment. For Fischer and Ravizza, an agent is morally responsible insofar as the mechanism that leads to action is moderately reasons-responsive. Just as endorsement theories err by construing endorsement in singular terms, control theories err by construing control in singular terms. The problem with Woody Allen agents is not that they fail to do what their best judgment tells them to do, or that they would not do otherwise in the face of countervailing reasons. The problem is that what they take themselves to have reason to do conflicts with their alienation from those reasons. (I consider whether Mele's discussion of self-control over emotion has the resources to address this tension in n.18 below.) Thanks to an anonymous reviewer for this journal for suggesting this extension to the scope of my critique.

[11] The theory is being developed in more detail in Killmister (MS)

[12] I'm using the term 'motivational attitude' broadly, to cover basic desires as well as more complex psychological states such as emotions, insofar as these incline the agent towards some kind of action. On this characterization, motivational attitudes necessarily involve a desire to act, but they do not reduce to those desires (c.f. Mele 1995b). For instance, both fear and disgust typically involve a desire to avoid their object, but this desire is coupled with very different representations of that object. This will prove to be important, as the explanation of an agent's deeming a motivational attitude unreasonable may only be explicable once we introduce her representation of its object as, i.e., disgusting or dangerous, over and above her desire to avoid it.

Furthermore, the term 'motivational attitude' is not intended to extend to beliefs, unless these beliefs encompass a desire to act. So a belief that an object is disgusting does not constitute a motivational attitude, unless it carries with it a conative response. This clarification is important to understand what is meant to be going on in the original twinkie case: while the agent believes the twinkie to be disgusting, she does not *experience* it as disgusting. The only motivational attitude she experiences is desire.

[13] See, for instance, (D'Arms and Jacobson 2000)

[14] It's also possible that the agent doesn't invoke any critical standard at all. We can imagine an agent with a particularly romantic bent, for whom the mere presence of a motivational attitude is sufficient to justify it as reasonable. This needs to be distinguished from the agent who tolerates her motivational attitudes being unreasonable. For such an agent a standard is still being invoked and deemed unsatisfied; the agent just remains unmoved by this fact. For the former kind of agent, all motivating attitudes are taken to be reasonable by default. For the latter agent, by contrast, many attitudes are taken to be unreasonable, but the agent remains unperturbed by this fact about herself. It is only in the case of the latter agent that there is a lowering of Self-Definition.

[15] This may raise worries about agents who are deceived or otherwise epistemically compromised. An agent may fail to recognise that an attitude is unreasonable because she has been indoctrinated, self-deceived, or because of some other cognitive impairment. Clearly we want to say that in cases such as these subjective endorsement is insufficient to secure the agent's autonomy. This points to the need to supplement actual endorsement with some kind of counterfactual endorsement. In the broader theory from which this paper draws I incorporate an idealisation – building on the work of John Christman, I argue that endorsement is only fully secured if it would persist in light of fuller knowledge conditions (the details of which are to be set by the agent's own epistemic commitments). Since *ex hypothesi* the Woody Allen cases aren't due to deception or any other cognitive impairment, and there's no suggestion that their judgment of unreasonableness would be retracted with fuller knowledge conditions, I leave the details of this aspect of the theory to one side here.

[16] This is a key, though subtle, way in which my theory differs from Al Mele's account of self-control. While Mele incorporates control over emotions as part of what it means to be maximally self-controlled, and thus as a necessary condition for full autonomy, he sees this in terms of the agent judging that she should not have a particular emotion, and then bringing about the removal of that emotion. To put it slightly differently, Mele's account of self-control over emotions parallels the demands of Internal Self-Realisation: does the agent (intend to) bring about in herself that which she takes herself to have most reason to bring about.

[17] I borrow the language of bracketing from Scanlon (1998, 33-37), although it should be noted that Scanlon speaks of 'considerations' being bracketed, rather than motivational attitudes. This view has interesting parallels with McDowell's (1979), notion of silencing, though divorced from the context of virtue.

[18] This may seem to beg the question against theories of autonomy which build in a strong responsiveness-to-reasons condition. Space limitations prevent a full defense of my theory here, but one key consideration against taking autonomy to have such a condition involves the explanatory power of the competing theories. Since any theory of autonomy is presumably going to be at least in part prescriptive, we need to ask what work a theory of autonomy should be able to do. I take it as a strong desideratum that a theory of autonomy should be able to make sense of the problem of paternalism. It's worth noting that theories with strong substantive conditions on the content of agent's desires will have trouble saying why interventions to prevent the fulfilment of 'deviant' desires are problematically paternalistic. In fact, such interventions will presumably not even be able to count as hard paternalism, since the action they are preventing is not autonomous.

[19] I must stress at this point that I do not take justifying reasons to be as restrictive as Bratman does. Given the subjectivity of ISR, the only restrictions on what can count as a reason for action are determined by the agent herself. This makes the theory much less demanding than Bratman's: insofar as an agent takes relief of the discomfort of a desire to be a reason, then acting to fulfil the desire doesn't in and of itself make her less autonomous. With this in mind, we can see one way in which the opening examples may differ. It is at least plausible that agents typically treat unreasonable desires as less reason-giving than unreasonable fears, since refusing to act on such desires typically carries less psychological burden than refusing to act on such fears.

[20] This is not to deny that for most agents, most of the time, the fact that an attitude is unreasonable will be sufficient to reject it as a reason for action. It may well be that a tight relationship between reasonableness and reasons for action is far more common than their coming apart. However, all that is required for my theory to be plausible is that *in at least some cases* the agent takes a motivational attitude to be both unreasonable, and a reason for action. In other words, my theory requires that it be possible to distinguish between these two attitudes, and plausible that some agents treat them differently.

[21] C.f. Sarah Buss (1994, 102), discussing the 'unwilling addict': "if we fail to distinguish between attitudes toward *having* a given desire and attitudes toward *being moved by* this desire under these very circumstances

– then we will think it obvious that people can be forced to intend to do something they prefer not to do, all things considered." It is worth noting that the first half of Buss' distinction is different to mine: as noted above, not wanting to have a desire is compatible with judging that desire to be reasonable.

[22] For an overview of the exhibit, see http://www.exploratorium.edu/tv/index.php?program=1062&project=58. As an anonymous referee for this journal has pointed out, many agents may find their attitudes of disgust in such cases to be entirely reasonable. I think this is probably right: may point here is simply that *even if* agents find such attitudes to be unreasonable, this does not typically result in those attitudes dissipating.

[23] I also have a tendency to weep at Adam Sandler movies, but that is a more explicable phenomenon.

[24] Moreover, such agents' akrasia would be further exacerbated by the fact that the motivation from which they act is one they take to be unreasonable (see the discussion of doubly-unwilling addicts below).

[25] I leave aside here the worry that Frankfurt's description of addiction may bear little resemblance to the actual phenomenon of addiction (c.f. Levy 2006). For my purposes, what matters is that addicts may have a complex attitude towards their desire for the drug, and how this relates to what they take themselves to have reason to do. Such complex attitudes, I hope, do not stray too far from the real-world experience of addiction.

[26] A couple of points about the autonomy of the willing addict are worth stressing. First, it's important for the claim that she is fully autonomous that the intention to take the drug is due to that being what she takes herself to have reason to do, rather than just fortuitously corresponding to it. Likewise, it's important that the action is brought about by the intention, rather than merely fortuitously corresponding to it. Insofar as either of these connections are weakened, the agent's autonomy is compromised.

[27] As an anonymous reviewer for this journal pointed out, we might also wonder what to say about the akratic non-addict. As s/he notes, on Frankfurt's model the akratic non-addict (who intentionally takes the drug while not identifying with her desire) would have to be construed as unfree and hence not responsible. On my theory, akrasia will primarily be captured by Internal Self-Realisation: insofar as akratic action consists in intending to do other than what we take ourselves to have most reason to do, then akratic agents will have low Internal Self-Realisation, and correspondingly reduced autonomy. (If the akrasia manifests between the intention and the action, by contrast, it will be captured by External Self-Realisation.) It is important to stress, however, that the judgment of lowered autonomy for the akratic non-addict would not necessarily correspond to a reduction in moral responsibility. On my broader theory, the degree of moral responsibility an agent holds for an act does not neatly correspond to the degree of her autonomy for that act. Very briefly, I take moral responsibility to be more sensitive to reductions in External Self-Realisation than reductions in Internal Self-Realisation. (In other words, failure to act as we intend makes us less morally responsible; failure to intend in accordance with what we take ourselves to have reason to do does not).

[28] I leave aside here the possibility of addicts who consume the drug despite never forming an intention to do so. Such agents, on my theory, would exhibit drastically compromised External Self-Realisation. With respect to such agents, I think that Sarah Buss (1994, 103) is right when she suggests we may sometimes form intentions to do that which we expect we will be unable to avoid doing, as a way of protecting our autonomy: "Resignation can be the attitude of last resort for someone who values nothing more dearly than her autonomy." Translated into the terms of my theory, such an agent would be protecting her External Self-Realisation at the expense of her Internal Self-Realisation.

[29] This may be the case for desires for some forms of self-harm, which the agent might reject as unfitting on the grounds that to fulfil them would bring her pain but no pleasure. (Of course, an agent with sadistic tendencies would find pleasure in the pain, and so if she were to reject her desires it would have to be on different grounds.)

[30] There is, of course, a third kind of unwilling addict, corresponding to the Woody Allen case. This would be an agent who took her desire for the drug to be unreasonable, but nonetheless took it to be a reason for action. It follows from my theory that both the Woody Allen addict and the akratic addict would, *ceteris paribus*, be more autonomous than the doubly unwilling addict. As an anonymous referee for this journal points out, this latter judgment may be resisted. That is, someone might claim that aligning one's judgments of reasonableness and reasons for action is itself autonomy-enhancing. Presented in that form, the objection seems correct. To take one's judgment of unreasonable to commit oneself to bracketing that attitude as a reason for action shows that the agent is extending her judgments throughout her deliberation, which is to exhibit an important component of self-governance. However, the plausibility of the objection is brought into serious doubt once it's built into the example that the agent *acts* from her unreasonable attitude. In such cases (of which the doubly-unwilling addict is one) we have an agent whose action is entirely divorced from her practical standpoint. The action fails to be endorsed in any way, and thus fails to be an expression of the agent's self. This leaves open which of the two following agents would be more autonomous: the agent with

an unreasonable desire, who bracketed that desire as a reason for action, and then did *not* act from it; or the Woody Allen agent, who has an unreasonable desire, but takes that desire up as a reason for action and then acts on it. I leave that puzzle aside for another day.

References

Bratman, Michael. (1999) *Faces of Intention: Selected Essays on Intention and Agency* (Cambridge: Cambridge University Press).

_____ (2007) *Structures of Agency; Essays* (Oxford: Oxford University Press)

Buss, Sarah. (1994) "Autonomy Reconsidered", *Midwest Studies in Philosophy* 19 (1): 95-121

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

_____ (2007) "Autonomy, History and the Subject of Justice", *Social Theory and Practice* 33 (1): 1-26

_____ (2009) *The Politics of Persons: Individual Autonomy and Socio-Historical Selves* (Cambridge: Cambridge University Press)

D'Arms, Justin and Daniel Jacobson. (2000) "The Moralistic Fallacy: On the 'Appropriateness' of Emotions", *Philosophical and Phenomenological Research* 61 (1): 65-90

Fischer, John Martin and Mark Ravizza. (1998) *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press)

Frankfurt, Harry. (1998) *The Importance of What We Care About* (New York; Cambridge University Press)

Killmister, Suzy. (MS) *Taking Autonomy's Measure*

Levy, Neil. (1996) “Autonomy and Addiction”, *Canadian Journal of Philosophy* 36 (3): 427-448

McDowell, John. (1979) “Virtue and Reason”, *The Monist* 62 (3): 331-350

Mele, Alfred R. (1995a) *Autonomous Agents: From Self-Control to Autonomy* (Oxford: Oxford University Press

_____ (1995b) “Motivation: Essentially Motivation-Constituting Attitudes”, *The Philosophical Review* 104 (3): 387-423

Plato. (1974) *The Republic* trans. Desmond Lee (Middlesex: Penguin Books)

Scanlon, Thomas M. (1998) *What We Owe to Each Other* (Cambridge, MA: Harvard University Press)

Velleman, J. David. (2000) *The Possibility of Practical Reason* (Oxford: Oxford University Press)

_____ (2007) “What Good is a Will?” in Anton Leist (ed), *Action in Context* (Berlin: De Gruyter)

Wolf, Susan. (1993) *Freedom Within Reason* (New York: Oxford University Press)