NOÛS 51:1 (2017) 3–23
doi: 10.1111/nous.12106

Agency and Inner Freedom

MICHAEL GARNETT
Birkbeck College, University of London

What is it to be rendered unfree, not by external obstacles, but by aspects of oneself? We can lack freedom in many ways: we can be shackled in chains or locked in cages; we can be stranded on islands or trapped in caves; we can be subjects of legal prohibitions, targets of social sanctions, or victims of private threats. In all of these cases we lack freedom by virtue of something external to us, something 'outer'. Yet we can also be driven by addictions, inhibited by phobias, or goaded by irrational compulsions; we can be limited by ignorance, blinded by prejudice, or trapped by superstition; we can be brainwashed by cult leaders, controlled by ideology, or subject to oppressive internalised norms. In many of these cases we also lack freedom, but by virtue of something internal to us, something 'inner'. The question is how to understand such losses of inner freedom.

This paper concerns the relationship between this question and another that is often taken to be intimately connected with it, namely that of how to understand failures of agency. What is it to fail at being a *doer*, a thing that genuinely acts, and instead to be a thing that is merely acted upon, passive in relation to its own behaviour? It is widely believed that the answer to this question about agency holds the key to the question about inner freedom. The thought is that what it is to be rendered unfree by an aspect of oneself is, at least in part, to be in some way alienated from or passive with respect to it, and for it to constitute an external force. For this reason, these questions about freedom and agency are often taken together, and the search for an account of inner freedom treated as equivalent to the search for an account of 'true' or 'deep' agency (Christman and Anderson 2005; Christman 2009; Frankfurt 1971; C. Taylor 1979; Velleman 2000, 2002; Watson 1975).

In this paper I argue that this is a mistake. Losses of inner freedom are not explicable in terms of failures of deep agency, that is, in terms of motivation by alien desires. This is surprising: in addition to overturning a widespread assumption of much contemporary work on agency and personal autonomy, this result also challenges what is, on at least one interpretation of the tradition, the doctrine at the heart of 'positive' approaches to political freedom, namely the idea that to be truly free one must be able to give expression to one's true self (C. Taylor 1979). Yet while this severing of the theory of freedom from the theory of agency may seem radical when presented starkly, it is in fact consonant with a more recent turning away, on the part of some autonomy theorists (Arpaly 2003; Garnett 2014; Mele 1995; J. S. Taylor 2009), from those questions about identification and

alienation that so animated philosophers in previous decades. If successful, this paper provides a deep vindication of this new emerging theoretical orientation.

I. Agency and Freedom: Three Views

Consider one of the central characters from many contemporary discussions of agency and inner freedom: the unwilling addict, whom we shall call Jane. Jane is not a real-world addict, but rather that oversimplified figment of the philosophical imagination that is supposed to serve as an analytic tool by modelling a wide range of cases.[1] All we need to know about Jane is this: she wishes she were not an alcoholic (say), and wants desperately to do various things incompatible with her alcoholism, but nevertheless is ultimately always moved by her urge to drink. I here leave aside the important and much-discussed issue of Jane's moral responsibility.[2] Instead I focus on a different issue. One common idea is that Jane lacks full inner freedom, that there is at least some sense in which her own psychological state, her alcoholic urge, renders her unfree. Another common idea is that Jane lacks full agency, that the urge that moves her is in some way alien to her true motivations. My question is in what way the first of these intuitions might rest on or be in part explained by the second. In this section I briefly introduce three different, though mutually consistent, ways in which this might be taken to be so; while these lines of thought are not always carefully distinguished from one other, all three ought to be very familiar.

The first begins with the idea that inner states can compromise freedom in essentially the same way as outer obstacles can, namely by *preventing* the agent from doing what she would otherwise be able to do. On this way of thinking, Jane's addiction undermines her freedom by constituting a constraint on the successful pursuit of her true purposes. This approach leads us to a view concerning the nature of Jane's agency because, in order for her addiction to count as a constraint, there must be some sense in which it is something that Jane *faces* or *confronts*, and therefore some sense in which it is alien or external to her true agency. As Alan Wertheimer puts it:

> ... the structure of freedom discourse always presupposes *some* distinction between the *agent* who is constrained and the *constraint* itself. There is a radical sense in which constraints are always *conceptually* external, even if they operate within the person. We can, for example, speak of addiction as an internal or psychological *constraint*, but only because we can posit a purified conception of the agent that is distinguishable from the 'empirical' conception that includes the addiction. (Wertheimer 1987: 261)[3]

So the idea of inner constraint necessitates a theory of deep agency, an account of the 'agent proper' that can be thought of as distinct from some of its desires. This is the first way in which claims about inner freedom might be taken to rest on claims about agency.

The second way proceeds, not via the idea of inner constraint, but via that of *self-government*. Self-government is, of course, an ideal of freedom: a society that lacks self-government is, in one recognised sense, unfree. Yet in what does self-government consist in the individual case? A natural thought here is that *self*-government is government *by the self*: in Laura Ekstrom's words, 'one's action is self-governed

when it is directed by the true self' (2005: 155). So Jane lacks self-government because Jane—the true Jane, who wishes to stay sober—is not adequately in charge of the wider Jane's behaviour. Thus, once again, we need a theory of true agency or selfhood in order to make sense of Jane's loss of inner freedom. This is the second way in which claims about inner freedom might be taken to rest on claims about agency.

The third way rests on the idea of *independence from external control*. When a person is subject to the unauthorised control of something else—such as the coercion of another agent—then that person lacks freedom. In this sense, inner freedom is a matter of acting under the control of forces genuinely internal to one's agency. Hence Jane is unfree because her behaviour is determined by alien motivations: rather than acting in her own right, she is a slave to her addiction. To make sense of this idea, we need a theory concerning what it is for motivations to be 'internal' or 'external' to agents: in short, a theory of deep agency. This is the third way in which claims about inner freedom might be taken to rest on claims about agency.

Each of these three thoughts invokes a conceptually distinct notion of freedom. The first relies on a broadly 'negative' conception, according to which freedom is the absence of constraint. The second relies on a 'positive' conception, according to which freedom is the presence of self-government. The third relies on what we might think of, in a very wide sense, as a 'republican' conception, according to which freedom is independence from external control (c.f. James 2009). Indeed, it is this that underlies the three approaches' mutual consistency. Thus Jane's addiction might compromise her inner freedom in all three senses at once: by preventing her from doing what she truly wants, by undermining her attempts to govern herself, and by constituting her subjection to an external force.

As just indicated, each of these claims seems to rest on a view about what it is for aspects of one's mind to be truly one's own. Elucidating this notion has been perhaps the central preoccupation of action theory over the past several decades, during which time a number of rival accounts have been developed (e.g. Buss 2012; Christman 2009; Ekstrom 2005; Frankfurt 1971, 1999; Korsgaard 2010; Moran 2001; Velleman 1992, 2002; Watson 1975). I am not, however, concerned with this problem, or with the merits of the different accounts, and I have nothing to add to these debates here. Instead, I am concerned with the more basic issue of the role that these accounts' shared *explanandum*—alienation from a desire—should play in our theory of freedom. Thus in this paper I simply *assume* that we have some settled account of what it is for a desire to be truly an agent's own, without getting into the details of what this account requires (readers may have in mind their preferred account). My aim is to argue that, *whatever* account we give of agency in this deep sense (whether involving higher-order endorsement or reasons-responsiveness, whether historical or relational or both, and so on), and however important that account is for other philosophical projects, it is of no help—indeed, it is a great hindrance—in the project of understanding what it is for an agent to be rendered unfree by aspects of herself.

Thus I argue that all three views about the relation between freedom and agency are problematic. If we want to understand Jane's loss of inner freedom in any of these ways, we must revise them so as to detach their associated claims

about agency. Hence if we want to think of Jane's addiction as a constraint, we must not think of constraints as essentially external to that which they constrain; if we want to think of Jane as lacking in self-government, we must not think of self-government as government by the (true) self; if we want to think of Jane as subject to a relevantly external force, we must not think of agents as things that are necessarily independent of relevantly external forces. We must allow clear water between our theory of inner freedom and our theory of deep agency.

I argue for these claims in turn: Sections II–IV concern constraint, Section V concerns self-government, and Section VI concerns independence. Section VII concludes the argument.

II. The Idea of Inner Constraint

I begin by further exploring the supposed connection between inner constraint and alienation: why, exactly, should a theory of the former require a theory of the latter?

The idea of *outer* constraint is relatively straightforward. It can be understood, in familiar terms, as essentially a matter of what an agent would fail to do were she moved to do it (e.g. Hobbes 1991: 146). So why should inner constraint be any more troublesome? To be sure, not all forms of inner constraint *are* more troublesome. Suppose, for instance, that a prison cell's door is fitted with an accessible combination lock; were the prisoner to know the correct combination, he could open the door and leave, but he does not. In this case, we might say that (part of) what prevents the prisoner from leaving the cell is his own ignorance, where this, being a fact about his belief set, is properly classed as a form of inner constraint. Such *non-motivational* inner constraints need pose theorists no special problems.

The special problems start, however, when we try to understand how agents can be constrained by their own motivations. For suppose that I am in a room with an open door, but that I do not want to leave it because I am happily reading a book (Locke 1975: 266). Does my desire to read *prevent* me from leaving the room, in the way that the prisoner's ignorance prevents him from leaving his cell? Certainly not: ordinary motivations cannot be treated as constraints, on pain of making a mockery of the very ideas of 'option' and 'constraint'. To treat motivations *per se* as constraining is to hold that one is always constrained from doing other than that which one is currently doing, and so to hold that one never enjoys more than a single option (that of doing what one is most motivated to do). Yet it is essential to our ordinary idea of an option that options need not be exercised: what a person *can* do, and what a person *will* do, are separate matters. Ordinary motivations cannot be constraints.

In light of this, return to Jane. Suppose that we naïvely treat Jane's urge to drink as just another one of her motivations, fundamentally no different from my desire to keep reading. Then we have no grounds on which to think of her urge as constraining, any more than we do my desire; and we are therefore committed to the view that Jane has the genuine option of staying sober, just as I have the genuine option of putting down my book. If we are coherently to view Jane's

alcoholic urge as a constraint, then, we cannot view it as just another one of her motivations. Instead, we need a theory that identifies some relevant property of Jane's urge—something that marks out *this* motivation as against her other motivations—and allows us to treat it as constraining without forcing us to the intolerable view that motivations are constraining *per se*.

This is why motivational constraints are especially problematic and demand a special theory. But why a special theory of *agency*? The thought is that a constraint, even a mental constraint, is essentially something that an agent *confronts*; it is 'a bit of reality for the agent to accommodate' (Moran 2001: 115), something that the agent may plan around, or attempt to overcome, but nevertheless something that is *faced*. So the special feature of Jane's addiction—what marks it out from her other desires as something that might function as a constraint—is its externality, the fact that she relates to it as something 'out there' and not 'in here' (c.f. C. Taylor 1979: 153–4).

Let us state the underlying principle on which this line of thought rests as follows, namely:

> *The Principle of Constraint Externality*: For C to be a constraint on A, there must be some relevant sense in which C is external to A.[4]

This is the principle that takes us from the claim that Jane is *constrained* by her alcoholism to the claim that it is somehow *external* to her.

If we accept Constraint Externality, and if we accept that agents can be constrained by their own motivations, then we are committed to accepting (what I call) a 'deep' view of agency. In discussing this it will be useful to distinguish between two models of agency. On the *basic model*, which we may think of as a Hobbesian or Humean model, the agent is straightforwardly identified with the totality of her motivations, desires and inclinations. This model permits many types of discrimination amongst motivations: they may be rational or irrational, reflective or unreflective, fleeting or enduring. What it does not permit, however, is discrimination amongst motivations in terms of the extent to which they are or are not truly the agent's own. The basic model has no room for this thought; it treats all motivations as *equally* the agent's own. This is a simple—most would say simplistic—and theoretically undemanding model of agency.

As we have seen, anyone who accepts both Constraint Externality and the possibility of motivational constraint must find a way to move beyond this basic model. They must recognise a deeper sense in which an agent may be alienated from her own motivations. On the simplest version of this 'deep agency' model—the *dual model*—it is recognised that the basic model does capture something correct about agency: after all, there is undeniably *some* sense in which Jane's alcoholic compulsion is her desire, and drinking her action. But these, according to the dual model, are merely shallow or wide senses of these terms. Thus there is, additionally, a deeper sense in which Jane's desire is not truly her own, and drinking not genuinely her action. The dual model therefore recognises both a *shallow* agent identical with the agent of the basic model, and a *deeper* agent identified just with some subset of the shallow agent's motivations.

In speaking here of 'deep agency' I have in mind only the familiar idea that there is some sense in which not everything that occurs in a person's mind is truly ascribable to her as an agent properly considered (c.f. Frankfurt 1988). As Richard Moran puts it:

> An obsessional thought that a person feels alienated from is nonetheless an episode in the psychological history of *that* person and no other... We may call this the weak or 'empirical' sense in which the thought or movement is his. But the fact that we can describe such a person as *alienated* from his obsessional thought, as if coming to him from without, shows that there is also a stronger sense in which such a thought may or may not be experienced as 'his own'. (2002: 190–1)

I have said already that my discussion does not turn on how, precisely, we go about drawing the border between 'internal' and 'external' attitudes. Yet nor, importantly, does it turn on how exactly we conceive of the nature of the distinction itself. That is, it does not matter to my argument whether we conceive it, for instance, as a *metaphysical* distinction between ontologically discrete entities, as a *perspectival* distinction between different ways of looking at one and the same entity, or as a merely *phenomenological* distinction between different ways of experiencing ourselves. Inasmuch as each of these approaches constitutes some version of the idea that there is an important distinction to be drawn between 'deep' and 'shallow' notions of agency, it falls under the broad theoretical scope of what I am calling 'the dual model'.

Unlike the basic model, any variation of the dual model permits us to say that Jane's addiction is external to her (*qua* deep agent). In so doing, it permits us to regard her addiction as a constraint without violating the Principle of Constraint Externality. Yet it makes for a poor theory of inner unfreedom because, as I shall now explain, it thereby forces us to *deny* that agents can be rendered unfree by their own motivations—exactly the phenomenon that such a theory is required to explain.

III. Inner Constraint on the Dual Model

The task of the previous section was to understand how an agent, such as Jane, could be constrained by her own motivations, and the dual model was supposed to assist in this. Yet it does not. On the dual model the agent can be thought of either shallowly or deeply. The shallow agent is the simple Hobbesian or Humean agent that is (shallowly) identified with all of its motivations: *qua* shallow agent, then, Jane's alcoholic urges are internal to her agency, and constrain her no more than ordinary desires ever constrain anyone. The deep agent, by contrast, is identified just with certain motivations, and is alienated from others: *qua* deep agent, then, Jane's alcoholic urges are constraints, but not *internal* constraints; understood in this way, they are external to her agency. So neither one of the two agents of the dual model is constrained by its own motivations. The shallow agent counts the addiction amongst its own motivations, but is not constrained by it; the deep agent is constrained by it, but does not count it amongst its own motivations.

Indeed, this much simply follows from Constraint Externality. A constraint cannot be fully internal to the agent that it constrains. On this view, an agent may be rendered unfree and constrained by *a* motivational state, but not by *her own* motivational state. For these reasons it may look as though this broadly 'negative' conception of freedom, with its framework of opportunity and constraint, may be fundamentally incapable of helping to deliver an account of inner unfreedom. Indeed, the approach is perhaps better understood as one that attempts to *explain away* the phenomenon, *reinterpreting* claims about inner constraint as claims about outer constraint, and thereby denying that there is any such thing, strictly speaking, as being rendered unfree by one's own motivational states.

Lest anyone think this revisionary consequence unremarkable, note just how out of line it is with common ways of thinking. Consider the following statement:

(1) Jane lacks freedom inasmuch as her alcoholism prevents her from holding down a job.

On the approach we are considering, *all statements of this form are false*. In reading a statement like this we must decide whether 'Jane' here refers to Jane *qua* shallow agent or to Jane *qua* deep agent. If we take 'Jane' to refer to the shallow agent, then the statement is false because Jane's alcoholism does not prevent her from doing anything (any more than my desire to read a book prevents me from leaving my room). If we take 'Jane' to refer to the deep agent, then the statement is false because Jane (*qua* deep agent) has no alcoholic urges. So it is false either way. But statements like this can surely be true.

What *is* true, on the dual model, is the following:

(2) Jane (*qua* deep agent) lacks freedom inasmuch as her (*qua* shallow agent) alcoholism prevents her (*qua* deep agent) from holding down a job.

Thus the original statement appears true only insofar as we equivocate on the referent of 'Jane'. It is because it is somehow oddly tempting to take 'Jane' to refer, incoherently, to both deep and shallow agents simultaneously that the dual approach is able to generate the illusion of an intuitive solution to the problem of inner freedom. On closer reflection, however, it provides no such solution.

Charles Taylor writes (1979: 160):

> ... our significant purposes can be frustrated by our own desires, and where these are sufficiently based on misappreciation, we consider them as not really ours, and experience them as fetters. A man's freedom can therefore be hemmed in by internal, motivational obstacles.

Yet statements like this must be rejected (at least as they stand), since they equivocate on the nature of the agent in question. It is only *qua* deep agent that a man's freedom can be hemmed in by motivations that are not really his, while it is only *qua* shallow agent that such motivations are internal to him. So Taylor has not in fact explained how one and the same agent can be fettered by his own desires.

Perhaps it will be objected that this line of argument assumes too literal or metaphysical an interpretation of our talk about desires being 'external' or 'alien'. Yet the argument is, in fact, neutral with respect to how we conceive the dual model's central distinction. Understood perspectivally, we may observe that there is one perspective from which Jane's addiction appears internal to her agency (the 'third-personal' or 'empirical' perspective), and another from which it appears as a constraint on her agency (the 'first-personal' or 'deliberative' perspective); the problem, however, is that there is no *single* perspective from which her addiction is both internal to her agency and a constraint on it (so long, that is, as we accept Constraint Externality). Understood phenomenologically, it is a corollary of Constraint Externality that one can *experience* something as constraining only insofar as one *experiences* it as alien or external, and this makes the idea of a genuine experience of inner constraint every bit as problematic as that of inner constraint itself, and for the same reasons. Thus the dual model cannot be made to yield a non-equivocal sense in which statements like (1) come out true, no matter how it is interpreted.

In light of these problems, it is perhaps worth recalling the considerations adduced above (Section II) as to why some version of the dual model is indeed necessary if we wish to treat Jane's addiction as a constraint at all. The root idea is that constraints must be somehow external to the agents they constrain (this is the Principle of Constraint Externality). It is this idea that draws us towards a 'deep' model of agency in the first place: without it, there is no obvious reason to involve the framework of identification and alienation in our theory of inner freedom at all. In turn, it is only by invoking a sense of 'agency' that excludes certain of the agent's motivations that it is possible to make theoretical sense of the idea that Jane's addiction is external to her agency. Thus the dual model is unavoidable insofar as we wish to treat Jane's addiction as a constraint, given Constraint Externality. Yet the problem we now face is that this same model also *prevents* us from treating Jane's addiction as an *inner* constraint.[5]

Given that some 'deep' view of agency seems necessary in order to treat Jane's addiction as a constraint, then, perhaps it will instead be objected that the dual model, as I have characterised it, is from the start an uncharitable rendering of that basic idea. In particular, some may think it overly crude to insist that all motivations be classed as either 'internal' or 'external' to the deep agent. Instead, a more subtle view might be that internality admits of degrees: that while some desires are absolutely central to my identity as an agent, others are more peripheral, and that while some desires are only somewhat foreign to me, others are strongly alien to my purposes. This certainly seems to be how it *feels*: not every motivation from which I feel distanced, such as an unwelcome pang of jealousy, has the strongly alien character of an addiction or a compulsion. Moreover, whereas some existing philosophical theories of agency do seem to require a rigid division between 'internal' and 'external' motivations (e.g. Watson 1975), many do not. On a higher-order endorsement theory (Dworkin 1970; Frankfurt 1971), for example, we might understand the degree of one's identification with (or alienation from) a desire in terms of the strength of one's higher-order endorsement (or repudiation) of that desire, and on a reasons-responsiveness theory (Moran 2001; Scanlon 2002) we might understand the degree

of one's identification with a motivation in terms of the degree to which it is responsive to relevant reasons. So it may be contended that a 'fuzzier' model of agency, such as this, could avoid the problems just raised in relation to the dual model.

In the next section I argue that, whatever its advantages as a model of agency, such an alternative does no better than the dual model in helping to explain how agents can be constrained by their own motivations.

IV. The Fuzzy Model

On the *fuzzy model*, one's motivations are one's own to varying degrees. So we may say that Jane's urge to drink is *less* her own than her desire to stay sober; in acting on it, she acts to a *lesser* degree than she otherwise would. Given this, suppose that Jane's alcoholic urge is only 20% her own (supposing anything like this is, of course, entirely ridiculous; nevertheless, the fiction will aid clarity).[6] The problem is that the proponent of Constraint Externality now owes us an account of just *how* external a thing must be in order to qualify as a constraint. Moreover, wherever this bar is set, we will find ourselves driven inexorably back to what is, in effect, a dual model. For instance, we might say that, being 80% alien to Jane's agency, the compulsion is a constraint, in that it prevents her from doing what she more authentically wants to do. But then we have two agents again: the 'more authentic' agent, who is constrained, and the 'less authentic' agent, who is not. Indeed, we *have* to draw this kind of internal boundary in order to make sense of the compulsion as a form of constraint, since otherwise there is no sense in which the compulsion prevents Jane from doing what she wants to do; after all, taking the drug is what she wants to do, albeit with 20% authenticity.

Yet do not constraints themselves, in addition to the internality and externality of desires, admit of degrees? Perhaps these problems can be avoided simply by allowing that Jane's addiction, being only 80% external to her agency, is only 80% a constraint. It is certainly true that we often speak in terms of partial constraint, describing obstacles as more or less constraining. Unfortunately, however, a plausible account of what we mean in these contexts provides little help for those who might wish to accommodate the notion of internal constraint on the fuzzy model, as I now explain.

Here follows a brief account of 'partial constraint'. Suppose that you are walking through some mud. You are partially constrained, by the mud, from taking a step forward: were you walking on dry land, you would be unconstrained, and were you trapped in quicksand, you would be wholly constrained. How might we understand statements such as these? *Taking a step forward* is an act-type, and a fairly general one at that. If we move down a level to somewhat more specific act descriptions, we find that the idea of partial constraint no longer has application: we find only certain more specific acts (such as taking a brisk step forward) which you are wholly constrained from performing, and others (such as taking a laborious step forward) which you are wholly unconstrained from performing. This suggests that the point of talk of partial constraint may be to convey that

some but not all of the more specific acts included under a general act description are subject to (non-partial) constraints.

To elaborate: walking on dry land, you are constrained from performing *none* of the normal range of more specific actions by which you might qualify as taking a step forward; hence you are unconstrained from taking a step forward. Trapped in quicksand, you are constrained from performing *all* of the more specific actions of this type; hence you are wholly constrained. Walking in mud, you are constrained from performing *some significant proportion* of this normal range of more specific actions; hence you are partially constrained. Talk of partial constraint enables us to continue speaking at relatively general levels of description without having to spell out all of the more specific acts that a person is constrained or unconstrained from performing.

The claim that a person's alcoholism partially constrains her from doing her job, then, is plausibly interpreted as the claim that her alcoholism prevents her from doing her job *in certain ways* (e.g. well or reliably) though not in others (e.g. badly or unreliably). However, if we are to make sense of the idea of inner constraint on the fuzzy model, this is not a possible way of interpreting the claim that a partially alien desire, such as Jane's alcoholism, constitutes a partial constraint on her holding down a job. To this purpose, the partiality of the constraint must go all the way down; for otherwise we are still stuck, at the more specific levels of description, with a non-fuzzy notion of constraint. But there is no obvious sense to be made of the idea that one might be partially constrained from acting under a description that is *absolutely* specific (that is, a description that admits of no further specification). Talk of partial constraint is a shortcut; at the most specific levels of description, one is either constrained or one is not.[7]

So, in the relevant sense, constraints cannot come in degrees; so, to make sense of the notion of inner constraint, the fuzzy model must include a threshold of externality above which motivations can be constraints and below which they cannot; so it must, in effect, reintroduce the deep agent / shallow agent distinction of the dual model. Hence the fuzzy model does no better than the dual model in helping us to make sense of the idea of inner constraint. The logic of constraint simply demands a harder boundary between internal motivations and external obstacles than the fuzzy model supplies.

I have been arguing that there is a fundamental tension between two intuitively appealing ideas. One is the idea that a constraint is essentially something that one *confronts* or *faces*, and that constraints are therefore necessarily external or 'other', in some sense, to the agents that face them. This is the Principle of Constraint Externality. The other is the idea that agents can be constrained, not just by parts of the external world, but also by their own motivational states. This is the idea of inner constraint. We have seen that the standard way of trying to reconcile these two ideas fails. Separating deeper from shallower forms of agency gives the illusion of reconciliation, insofar as we fail to demand clarity concerning which of these, exactly, is the subject of our ascriptions of inner freedom and unfreedom. Yet no reconciliation has been achieved.

So we have two options. One is to deny Constraint Externality, and to find a way of treating Jane's addiction as both a constraint and genuinely her own. This means developing a new theory about what makes certain desires constraining—for if it is not their externality, it must be something else (since desires *per se* are not constraining); this might be, for instance, the desire's abnormal strength (Watson 1977), or its unresponsiveness to reasons (Glover 1970: 97–101). I do not pursue this project here. The relevant point is that, whatever the relevant property, it must *not* be one that entails that the desire is external to the agent whose freedom we are assessing, on pain of resurrecting the very problems we are here just putting to rest.[8] The result is that, if we abandon Constraint Externality, we have no special reason to move beyond the basic model—and, indeed, we have positive reason not to—in our attempts to understand inner constraint. Our theory of inner freedom no longer rests on a theory of deep agency.

The other option is to retain Constraint Externality and to conclude that this broadly 'negative' understanding of freedom as non-prevention is simply the wrong notion with which to try to make sense of the idea of inner freedom. This means, in effect, committing ourselves to the Hobbesian view (1991: 91) that, insofar as freedom is to be understood in terms of an absence of impediments, it must be understood only in terms of an absence of *external* impediments. This does not require, of course, that we agree with Hobbes and his followers about what counts as relevantly 'external'. But it does require us to concede that, however we draw the border between the 'inner' and the 'outer', freedom as non-prevention always attends to the latter. And this means, in turn, that in order to understand the idea of inner freedom we need to be working with a different notion of freedom.

V. The Self of Self-Government

I turn now to the idea that agents like Jane lack freedom, not in virtue of being prevented from acting, but in virtue of lacking self-government. Self-government is originally a political ideal pertaining to the internal organisation of groups or polities. The suggestion is that there is an analogous personal ideal—also worthy of the name 'self-government' and also, therefore, an ideal of freedom—that pertains to the internal organisation of individual agents. The question is how to understand it.

Let us first distinguish between 'negative' and 'positive' conceptions of self-government. In the 'negative' sense, self-government is simply a matter of being ungoverned by anyone else; in this sense, any independent sovereign nation is self-governing, regardless of its form of internal organisation. Understood like this, 'self-government' is just another way of talking about freedom in the sense of independence from external control, which will be discussed in Section VI. This section concerns the idea of self-government in its 'positive' sense.

The nature of positive self-government is a vexed issue in both the political and the personal cases. Here I want to focus on just one way in which the personal case is sometimes approached. This is the idea that *self*-government must be, in some sense, government *by the self*. David Velleman, commenting on cases like that of Jane, writes that 'such cases suggest that being autonomous, or self-governed,

is a matter of being governed from within—that is, by motives internal to the self' (2002: 92). Thus Jane lacks self-government because her self is not fully in charge of what she does. Velleman goes on to talk about the theoretical problem of identifying this 'self of self-governance' (2002: 93), a problem that he takes to be equivalent to that of developing an account of what it is for a motivation to be fully one's own (see also Ekstrom 2005: 155; Frankfurt 2002: 293; Noggle 1995: 57). The idea, then, is that any account of self-government must rely on an account of deep agency, since self-government just is government by the (true) self.

This familiar line of thought rests on something like the following principle:

> *The Principle of Self Government*: An agent is self-governed when she is governed by her true self.

If true, it immediately follows from this principle that we need something like the dual model of agency in order to understand self-government. However, the principle represents a confusion and is not, in fact, an explanation of what it is to be self-governed (Garnett 2011). The key is the hyphen. Self-government (together with self-rule, self-control, self-determination, self-legislation, and so on) is a reflexive relation—it denotes a relation that holds between a thing and itself. Moreover, a relation between a thing and *itself* is very different from a relation between a thing and its *self* (Cohen 1995: 68–9; O'Neill 2003: 16–7). There is nothing in the bare idea of self-government that invokes the idea of a self, much less that of a true or authentic self. Indeed, self-government no more invokes the notion of a self than does any other reflexive relation, such as self-fertilisation: a self-fertilising plant is not one that fertilises its self (that would be extremely difficult), but rather one that, simply, fertilises itself. By the same token, a case of self-government is simply one in which the thing doing the governing and the thing being governed are one and the same.

In light of this, the principle must be amended to:

> *The Principle of Self-Government*: An agent is self-governed when she governs herself.

The problem now, however, is that there is no longer any obvious reason to think that self-government requires government by an authentic self. Moreover, there is in fact positive reason to resist the thought, so long as we wish to understand losses of inner freedom as failures of self-government.

To see why, return to Jane and to the dual model. *Qua* deep agent, Jane wholeheartedly wants to stay sober, though she is prevented from doing so by forces external to her agency. Yet prevention by external forces is not a failure of self-government: external forces might prevent me from reading my book (the lights might go out), but that does not mean that I fail to govern myself. It is power, not self-government, that agents lack in virtue of failing to govern their external surroundings.[9] Of course, the 'external surroundings' over which deep agents may fail to govern include elements that are, in a different sense—the shallow sense—also parts of the agent; but to understand failures of self-government we need to understand what it is for one agent to fail to govern *that very same agent* (without equivocating on 'agent'). Hence the idea of deep agency is ultimately of little help in our attempts to understand failures of self-government because,

whatever else deep agents may fail to govern, they do not typically fail to govern *themselves*, and self-government is a reflexive relation.

Consider, in this connection, the well-known (albeit misplaced) objection to Kant's theory of autonomy that it ends up rendering heteronomous action impossible. According to this criticism, heteronomous behaviour is for Kant always the mere upshot of external forces operative on the will, and so never an agent's *doing*. As a result, such behaviour—being the mere effect of motivational forces at work within one—can never be genuine action. Thus pieces of immoral or irrational behaviour can never be imputed to agents, properly speaking. For Kant, therefore, it is alleged that there is no such thing as a heteronomous action and no such thing as an agent's failing to govern herself. What is important about this familiar problem in the current context is that it arises because, on this reading of Kant's position, autonomy or self-government is taken to be necessary for agency, and it is precisely this yoking together of autonomy and agency that rules out the possibility of heteronomous agency.

Of course, as is now well understood, this objection misconstrues Kant's position. For Kant, an event is attributable to my *agency* if it results from an exercise of my capacity for spontaneous choice (*Willkür*), where this exercise may or may not also be in accordance with the dictates of pure practical reason (*Wille*) and hence *free* in the sense of being autonomous. The result is that immoral or irrational behaviour can indeed be imputed to agents, so long as it is motivated by inclinations that have been 'taken up' into maxims in the appropriate way, as opposed to being causally determined by them (Allison 1990: 129–36). So for Kant agency is one thing, autonomous agency another. Kant avoids the familiar objection precisely by disassociating his theory of self-government from his theory of agential ascription in the manner that this paper advocates.

With this in mind, return once more to Jane and to the dual model. It is tempting to say something such as the following:

(3) When Jane drinks she is not self-governing, inasmuch as she is moved by an external force (her addiction) and not by herself.

Yet the present point is that this is, exactly, something that we cannot say. *Qua* shallow agent, recall, Jane's addiction is not an external force. *Qua* deep agent, we face a version of the anti-Kantian worry, for in this sense Jane does not act at all; she is a passive bystander to the external force that brings about her behaviour. This is not, then, a case of un-self-governed action, for it is not (in this deep sense) an action at all. And while Jane fails to govern the behaviour of her wider self, this is, *ex hypothesi*, simply a failure to govern some part of the external world, and we do not, normally, treat such external failures of government as failures of *self*-government. The lesson, again, is that when we line up our theory of autonomy too closely with our theory of agency, we render heteronomous agency impossible.

It may help to clarify these issues still further by taking a lead from Christine Korsgaard (2010) and considering Jane's moral psychology through the lens of Plato's city-soul analogy. Jane, as we have imagined her, strives to remain sober

but is routinely defeated by the sheer strength of her addiction. We might therefore think of her as like a city that finds itself powerless to stop some group from acting in persistent violation of its constitution, such as a group of fishermen who constantly break the city's treaty obligations by fishing in foreign waters without the constitutional authority to do so. Moreover, it is very natural to think of this latter, political case as one of failed self-government, and so it may seem that we have reason to view the former, personal case as one of failed self-government too.

Yet note that the thought that the city is failing at self-government depends crucially on the idea that the fishermen in some sense *belong* to the city, for without it we lose the idea that the city's problem is one of *self*-government. Suppose, for instance, that the city were to issue some similarly ineffective commands to members of a foreign city. This would not be a failure of self-government: since these foreigners are in no sense part of the city, in failing to govern *them* the city does not fail to govern *itself*. To secure the conclusion that the city's failure to govern the fishermen's activity is a failure to govern its own activity, then, we need some grounds for thinking of the fishermen as a genuine part of the city.

Exactly analogous considerations apply to Jane. In order to see her failure to control her alcoholism as a failure to control herself, we must treat that alcoholism as a part of her. Thus if we are to allow that there is such a thing as an agent's failing to govern herself, we must not treat her ungovernable elements as 'alien' or 'external' to her agency. This is because she fails to govern *herself* only insofar as she fails to govern elements that are *internal* to her agency. So not only does adoption of a deep agency model provide no assistance in our attempts to make sense of failures of self-government, it actively undermines them.[10]

A dual model of agency is unlikely to help us to understand a relation that holds between a single agent and itself. If we want to understand losses of inner freedom in terms of failures of individual self-government, therefore, we must do so with reference only to a basic model of agency.[11] Moreover, there are a number of approaches of this type: Alfred Mele (1995), for instance, gives a detailed account of self-control that eschews all appeal to 'true' selves (c.f. 1995: 123–6), and Onora O'Neill (2003) suggests a Kantian approach that does the same.[12] There is no need here to discuss these views, since my aim is not to determine the best account of individual self-government. It is simply to demonstrate that, if we are to understand losses of inner freedom in terms of failures of self-government, then, again, we must have clear water between our theory of inner freedom and our theory of deep agency.

VI. Agency and Inner Dependence

We come now to the last of the three ways of thinking about inner freedom. On this view, agents such as Jane lack freedom not because they face constraints, and not because they lack self-government, but because they lack *independence*: their behaviour is determined by external forces.

In thinking about what it is to be controlled by an external force, it will be helpful to borrow a version of Robert Kane's distinction between *constraining* and *non-constraining* control (1985: 33–4); that is, between forms of control that

essentially involve the removal of options and those that do not. This allows us to see that, insofar as 'independence from external control' refers to *constraining* control, it is more or less just a different, albeit more picturesque, way of talking about freedom in the broadly 'negative' sense of absence of constraint that has already been discussed (Sections II–IV). So if 'independence from external control' is to pick out a distinct conception of freedom, the relevant type of control must be non-constraining: we need to know what it is, not for one's purposes to be *blocked* by external forces, but for one's purposes *themselves to be the conduits of* external forces.

A natural thought is that what it is for a purpose to be the *conduit* of an external force is just for that purpose itself to *be* an external force. For after all, so this thought goes, to the extent to which behavioural events are the results of forces external to you, they are not really your doings at all; insofar as you are subject to external forces, that is, you are not functioning as a genuine agent, but rather as a kind of marionette with someone (or something) else pulling your strings. To be an agent is to be, in *some* sense, an originator of action in your own right,[13] and when you are subject to certain relevant kinds of external determination you are to that extent not an originator of action and so not an agent. Hence we cannot think of agents as fully active with respect to desires that are simply manifestations of external control, and we must think of them as alienated from desires that are the conduits of relevantly external forces. We therefore need a deep model of agency—that is, an understanding of what it is to be driven by purposes that are not fully one's own—in order to understand losses of inner freedom in terms of determination by external forces.

The underlying principle on which the preceding line of thought rests is something like the following:

> *The Principle of Agency Independence*: If, in acting on a motivation M, an agent A is subject to determination by external forces, then M is not truly a part of A's agency.

This principle connects the phenomenon of inner unfreedom—understood as subjection to external control—with that of deep agency. Thus suppose that you hypnotically implant in me, against my will, a desire to eat a lemon. In being motivated by this desire, we may plausibly say, I am subject to an external force (and hence unfree). Given Agency Independence, insofar as my being motivated by this desire constitutes my subjection to an external force, I am not functioning as a genuine agent, but am instead being passively moved by forces that are external to me. So, by this principle, we move from the idea that I am rendered unfree by my desire to the idea that I am alienated from it.

Hence it appears, again, that a deep model of agency is indispensable for an understanding of what it is to suffer a loss of inner freedom. Again, however, the appearance is misleading. The deep agency models are of no help in this regard—and for reasons that are by now familiar, though still worthy of clarification. Return, for sake of familiarity, to Jane and to the dual model. *Qua* shallow agent, Jane is determined by her addiction, but it is internal to her agency and so not an external force. *Qua* deep agent, Jane's addiction is an external force, but not one that determines her. Of course, it *constrains* her, and in *this* sense she is 'subject to' an external force; but in the 'non-constraining' sense, Jane is independent of

her addiction's control: *qua* deep agent, it does not drive her. So neither agent is determined by an external force: the dual model gets us nowhere.

Once again, the fuzzy model may seem more promising. For perhaps we can say that, if Jane's addiction is 80% alien to her (to resume the earlier fiction), then in drinking she is 80% subject to external forces; so the more alien a desire, the greater the extent to which the agent is determined by an external force and the less free she is. However, matters are not so simple. On the fuzzy model, Jane is only 20% truly involved in the behaviour that is her drinking. If it is *Jane's* freedom that we are interested in, then, we must look only at the nature of *Jane's* activity. And to the 20% extent that her drinking manifests her own activity, that activity may well be 100% independent of external forces.

An analogy may be helpful. Suppose that I am pushing a car along a road with four others, and that I contribute 20% of what is required to move the car. We might say that, of the process that is 'the pushing of the car along the road', 20% is constituted by my activity. Suppose now that we are interested in the extent of my freedom, in the sense of my independence from external forces. What we will look at, then, is the extent to which *my* activity—which represents only 20% of the total activity of pushing the car—is independent from external forces. We will not say that, since the pushing of the car is 80% the result of forces external to me, I am to that extent unfree or subject to external forces. *My* activity may be perfectly independent. In the same way, if we are interested in assessing Jane's independence from external forces, we need to look at the extent to which *her* activity—which represents only 20% of the total activity of her drinking—is independent from external forces. And, for all that we have said, there is no reason to think that this is not perfectly independent. So the fuzzy model gives us no reason to think that Jane is subject to relevantly external forces.

The root of this problem is the tight way in which Agency Independence again binds the idea of inner freedom to that of agency. Indeed, as may already be clear, it in fact follows trivially from Agency Independence that, necessarily, no agent can be determined by external forces (and that, therefore, no agent can suffer a loss of freedom in this sense). To see this consider that, in the case in which you hypnotise me into eating a lemon, we might naturally wish to claim something like the following:

(4) I am unfree inasmuch as, in eating the lemon, my action is determined by an external force.

Yet Agency Independence prevents us from claiming this, since by that principle *all statements of this form are false*. According to that principle, if eating the lemon results from external determination, then it is not truly *my action* at all; so I am not, after all, determined by an external force; so I am not, after all, unfree. Given Agency Independence, there can be no such thing as an agent whose actions are determined by external forces. So, if we are to understand inner freedom in terms of independence from such external determination, then we have no choice but to reject Agency Independence.

Note that this is a genuine 'if': I am not here denying the obvious appeal of Agency Independence. Among other things, the principle plays an important role in the arguments of those incompatibilists, such as Helen Steward (2012), who hold that determinism rules out agency itself, on the grounds that if all of our behaviour is subject to external control then we do not really *do* anything at all. My present point is just that, if we take a position such as this, then the idea of freedom as independence from external control is useless for understanding what it is for an agent to suffer a loss of inner freedom—since if we take a position such as this, we rule out the very idea of *unfree agency* ('unfree' in the sense of subject to external control).

By contrast, if we reject Agency Independence then we give up the reason we originally had for insisting that motivations that constitute agents' subjection to external forces must themselves be external forces, and we thereby sever the apparent conceptual link between inner unfreedom (understood as subjection to external control) and failures of deep agency. In our attempts to understand inner freedom in terms of independence from external forces, then, it seems we have no reason to move beyond a basic model of agency. Moreover, we have positive reason not to move beyond a basic model, for in treating independence-undermining motivations as external to the agent we lose the idea that it is the *agent's* independence that the motivations undermine. In order to avoid this problem, we must avoid treating the motivations by means of which agents are subjected to external forces as themselves external forces, that is, as motivations from which the agents in question are alienated.

Yet how are we to make sense of the idea of subjection to external control *without* appeal to the idea of alienation from one's desires? The key lies in locating the relevant externality not in the motivation that is now moving the agent to action, but in that motivation's causal history.[14] When you hypnotise me into eating a lemon, you subject me to an external force. In order to understand this, however, we do not need to suppose that my newfound desire to eat a lemon is *itself* an external force—it may be a thoroughly internal force. What matters is that *you* are, in this scenario, a relevantly external force, and that in eating the lemon I am subject to *your* control. The relevant relation of externality is between me and you, not between me and my desire. For while my desire is the *conduit* of an external force (your will), and while motivation by this desire *helps to constitute* my subjection to an external force (your will), there is no reason for us to treat my desire as *itself* an external force. Indeed, there is (as we have just seen) positive reason not to, since this undermines the idea that it is after all *me* who is subject to an external force in the first place.

Moreover, this basic lesson applies to any attempt to analyse freedom in terms of independence from external control. Some incompatibilists, for instance, may regard an agent as subject to external control, and hence unfree, if her action is causally determined by the distant past. Yet, again, the external force to which the agent is here subject must be the *past* (or determinism) and not her own determined desire, which may be perfectly internal to her agency (indeed, it must be internal to her agency, for else *she* is not causally determined by the distant past). However we understand 'external control', then, we must be careful not to suppose that externality is transitive, so to speak, from the causal origins of the

desire to the desire itself. Agents can be unfree by virtue of having desires with certain causal origins, even if those desires are entirely internal to their agency.

This is, in outline, how freedom from external control can be understood on a basic model, without appeal to deep agency. As I have argued in this section, if we are to try to understand losses of inner freedom in terms of subjection to external control, then we have decisive reasons for adopting this type of approach—that is, once more, for adopting an approach to inner freedom that does not rely on a theory of deep agency.

VII. The Impurity of Agency

What is it to be rendered unfree, not by external obstacles, but by aspects of oneself? Perhaps it is to be *constrained* by aspects of oneself. If we are to allow that one might be constrained by aspects of *oneself*, however, we must abandon the idea that constraints are necessarily external to those whom they constrain, and develop an account of inner constraint consistent with a basic model of agency. Perhaps, alternatively, it is to be incapable of *governing oneself*. If we are not to misunderstand this as an incapacity to be governed by one's *self*, however, we must take seriously its reflexivity, and develop an account of self-government that is, again, consistent with a basic model of agency. Perhaps, finally, it is for motivation by aspects of oneself to constitute one's *subjection to external forces*. If we are to allow that one's subjection to external forces might be constituted by aspects of *oneself*, however, we must reject the idea that agents are necessarily independent of such forces, and develop an account of non-subjection consistent with a basic model of agency. *However* we understand inner freedom, therefore, we need not—and, indeed, we must not—move beyond the basic model to a deep agency model.

The result is that we must be willing to see agency as less inherently *pure*. Agents themselves, and not just their external circumstances (including their mental circumstances), may involve elements that are controlled by outside forces, or are constraining, or are resistant to attempts at self-government. While the deep agency models appear to allow for this—indeed, while they appear necessary for us to make sense of this—they actually prevent us from accepting it, constantly hiving off the problematic items from the agent itself (the 'deep' agent) and relegating them to its external circumstances, helping to preserve the idea of a pure, free, uncompromised agent at the bottom of everything. This, however, is in direct tension with the common-sense idea that agents may be rendered unfree by aspects of themselves.

In recognising the potential impurity of our agency, we separate questions of agency from questions of freedom. This liberates us in two directions. As theorists of freedom, we can focus on developing our accounts of opportunity, self-government and independence without having to worry about questions concerning identification and alienation. As theorists of agency, we can focus on developing our accounts of identification and authorship without having to worry about questions concerning freedom or self-government. Once we allow that agents can be rendered unfree even by motivations with which they fully identify, these issues come apart.

I have been arguing that deep agency, and the theories of identification and alienation intended to elucidate it, are irrelevant to the problem of developing an account

of inner freedom. This is not to deny that these theories of agency are of crucial importance to other philosophical projects; to the contrary, there can perhaps be no final understanding of moral responsibility, or of the mind's place in nature, or of political power, or indeed of ourselves, until we understand what it is to be a genuine *doer* and not a mere locus of forces with respect to which one is ultimately passive. If the arguments of this paper are correct, however, we can nevertheless know what it is to be *free* in advance of this, with only a minimal and basic conception of agency.[15]

Notes

[1] For more nuanced and realistic philosophical discussions of addiction, see Elster (1999) and Elster and Skog (1999).

[2] For a recent illuminating discussion of this topic, see Watson (2004: Chs. 2, 3 & 11).

[3] See also Wright Neely (1974: 42–3): 'if we consider a desire as restricting an agent's freedom, we thereby relegate the desire to the circumstances of the agent's action and thus make a distinction between the agent proper and the desire'.

[4] That is, *at some particular time*: Jane might experience her desire to drink as a constraint at one time (e.g. when sober), and as 'internal' at another (e.g. at the bar); but at any one time she must experience it as *either* a constraint and so 'external' *or* 'internal' and so not a constraint, according to this principle.

[5] To illustrate this further, suppose that one were to seek to bypass the dual model by understanding Jane's relation to her addiction just in terms of a distinction between different ways in which motivations can belong to their agents (thanks to an anonymous referee at *Noûs* for raising this possibility). If Jane leases her apartment but owns her car outright, then there are two different senses in which these things belong to Jane, but only one agent to whom they belong; similarly, it could perhaps be argued that Jane's addiction and her desire to stay sober belong to one and the same agent, but in importantly different senses of 'belong'. However, if this move is intended to help explain how Jane can be *constrained* by her addiction, then (given Constraint Externality) it must be read as lending support to the idea that Jane's addiction is in some way *external* to her agency. And in order to see Jane's addiction as somehow external to her agency, we must have some ('deep') conception of Jane's agency that excludes her addiction—bringing us back, in the end, to some version of the dual model. On the other hand, if the move is not intended to support a claim of externality, then we lose the purported connection with inner freedom: for if there is no sense in which Jane's addiction is alien or external to her—if it is, in the end, entirely internal to her agency—then there is no obvious way in which it constrains her.

[6] Note that this does not mean, as in an ordinary case of mixed feelings, that drinking is 20% something that she wants to do and 80% something that she does not want to do. Nor does it mean that only 20% of the motivation is hers, as though the motivation were itself divisible. Rather, the motivation is 20% hers.

[7] See also Carter (1999: 223–34). Of course, I cannot rule out the possibility that there might be some better account of partial constraint that is more helpful to the fuzzy model in this regard, since I have shown only that one plausible analysis does not provide any help. Nevertheless this is, I take it, sufficient to hand over the burden of proof.

[8] If we are assessing the freedom of the shallow agent, could not the relevant property turn out to be externality-to-the-deep-agent? It could, although the invocation of externality is surely now superfluous. That is, suppose that one deems a (shallow) agent to be constrained by its irrational desires: even if one also happens to take irrational desires to be necessarily external to deep agents, it is the irrationality itself, and not the externality, that is doing the explanatory work, once Constraint Externality has been abandoned.

[9] Again, I am strictly agnostic about where 'the agent' ends and 'its surroundings' begin.

[10] It is unclear whether Korsgaard herself would wish to understand Jane's case in the way presented. *Were* she to accept this reading (c.f. 2010: 134–48), note that her theory would fail to explain how Jane suffers any loss of self-government in acting on her addiction, for exactly the sorts of reasons just discussed. This is because, for Korsgaard, the fact that a force is operating without constitutional authority

entails that its operation is not attributable to the person or city whose constitution it is (2010: 142). So since Jane's addiction operates without constitutional authority, it is not attributable to her agency; and in failing to govern it, therefore, she does not fail to govern *herself*. For this reason, Korsgaard might prefer a different reading of Jane's case, according to which Jane is not a well-constituted agent subject to extra-constitutional forces, but a badly-constituted one (c.f. 2010: 159–76). On this analysis, Jane's addiction is seen as a force operating *with* constitutional authority (albeit that of a 'defective' constitution), and as fully attributable to her agency. This avoids the problem just mentioned (albeit at the cost of regarding an unwilling addict as fully identified with her addiction, this being something of a departure from common assumptions of recent action theory). Importantly, in adopting this analysis of Jane's case Korsgaard would avoid the problem precisely by separating her theory of agential attribution (for her, a matter of constitutional alignment) from her theory of autonomy (for her, a matter of constitutional quality), i.e. by adopting a version of the position for which I am currently arguing.

[11] Granted that (reflexive) self-government is distinct from government by the self, and that the former is best understood on a basic model of agency, what entitles me to the assumption that it is indeed the former and not the latter that is relevant to my broader topic of inner freedom? What, after all, is in a word (or a hyphen)? The answer is that, in the current dialectical context, only self-government can be assumed without argument to be a type of freedom. For while the identification of self-government with freedom draws on a well-established political ideal, the same cannot be said for government by the self; this latter idea has no obvious political analogue and, insofar as one can be discerned, it is highly controversial whether the ideal it describes is indeed one of freedom (Berlin 1969). Thus the claim that government by the self is a form of freedom requires real argument. Moreover, it is precisely the thesis of this paper that government by the self does not in fact constitute freedom in any of its three uncontroversial senses (absence of constraint, self-government, and independence from external control).

[12] As does, perhaps, Korsgaard (2010)—see note 10.

[13] The qualifier is intended to encompass both compatibilist and incompatibilist conceptions of origination.

[14] As does, for instance, Mele (1995), though he does not motivate his approach by means of the present considerations.

[15] Many thanks to my colleagues Hallvard Lillehammer, Jennifer Hornsby and Susan James for valuable comments on an earlier draft. Thank you also to audiences at a 2013 Work-in-Progress Seminar at the Birkbeck College Philosophy Department and at a 2014 Colloquium at the Dalhousie University Philosophy Department (and especially to Joseph Millum for thoughtful discussion following the latter). Finally, thank you to an anonymous referee at *Noûs* for helpful comments on a previous draft.

References

Allison, Henry E. (1990), *Kant's Theory of Freedom* (Cambridge: Cambridge University Press).

Arpaly, Nomy (2003), *Unprincipled Virtue: An Enquiry Into Moral Agency* (New York: Oxford University Press).

Berlin, Isaiah (1969), 'Two Concepts of Liberty', in Isaiah Berlin (ed.), *Four Essays on Liberty* (London: Oxford University Press).

Buss, Sarah (2012), 'Autonomous Action: Self-Determination in the Passive Mode', *Ethics*, 122 (4), 647–91.

Carter, Ian (1999), *A Measure of Freedom* (Oxford: Oxford University Press).

Christman, John (2009), *The Politics of Persons: Individual Autonomy and Socio-Historical Selves* (Cambridge: Cambridge University Press).

Christman, John and Anderson, Joel (eds.) (2005), *Autonomy and the Challenges to Liberalism: New Essays* (Cambridge University Press).

Cohen, G. A. (1995), *Self-Ownership, Freedom and Equality* (Cambridge: Maison des sciences de l'Homme and Cambridge University Press).

Dworkin, Gerald (1970), 'Acting Freely', *Noûs*, 4 (4), 367–83.

Ekstrom, Laura (2005), 'Autonomy and Personal Integration', in James Stacey Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy* (New York: Cambridge University Press).

Elster, Jon (ed.), (1999), *Addiction: Entries and Exits* (New York: Russell Sage Foundation).
Elster, Jon and Skog, Ole-Jorgen (eds.) (1999), *Getting Hooked: Rationality and Addiction* (Cambridge: Cambridge University Press).
Frankfurt, Harry G. (1971), 'Freedom of Will and the Concept of the Person', *Journal of Philosophy*, 68 (1), 5–20.
——— (1988), 'Identification and Externality', *The Importance of What We Care About* (Cambridge: Cambridge University Press), 58–68.
——— (1999), 'On the Faintest Passion', in Harry G. Frankfurt (ed.), *Necessity, Volition and Love* (Cambridge: Cambridge University Press).
——— (2002), 'Reply to Jonathan Lear', in S. Buss and L. Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge, M.A.: Massachusetts Institute of Technology Press).
Garnett, Michael (2011), 'Taking the Self out of Self-Rule', *Ethical Theory and Moral Practice*, 16 (1), 21–33.
——— (2014), 'The Autonomous Life: A Pure Social View', *Australasian Journal of Philosophy*, 92 (1), 143–58.
Glover, Jonathan (1970), *Responsibility* (London: Routledge & Kegan Paul Ltd).
Hobbes, Thomas (1991), *Leviathan*, Richard Tuck (ed.), (Cambridge: Cambridge University Press).
James, Susan (2009), 'Freedom, Slavery, and the Passions', in Olli Koistinen (ed.), *The Cambridge Companion to Spinoza's Ethics* (Cambridge: Cambridge University Press), 223–41.
Kane, Robert (1985), *Free Will and Values* (Albany: University of New York Press).
Korsgaard, Christine M. (2010), *Self-Constitution: Agency, Identity, and Integrity* (Oxford: Oxford University Press).
Locke, John (1975), *An Essay Concerning Human Understanding*, Peter H. Nidditch (ed.), (Oxford: Clarendon Press).
Mele, Alfred R. (1995), *Autonomous Agents: From Self-Control to Autonomy* (New York and Oxford: Oxford University Press).
Moran, Richard (2001), *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton, N.J.: Princeton University Press).
——— (2002), 'Frankfurt on Identification: Ambiguities of Activity in Mental Life', in S. Buss and L. Overton (eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge, M.A.: Massachusetts Institute of Technology Press).
Neely, Wright (1974), 'Freedom and Desire', *The Philosophical Review*,83, 32–54.
Noggle, Robert (1995), 'Autonomy, Value, and Conditioned Desire', *American Philosophical Quarterly*, 32 (1), 57–69.
O'Neill, Onora (2003), 'Autonomy: The Emperor's New Clothes', *Proceedings of the Aristotelian Society, Supplementary Volumes*, 77, 1–21.
Scanlon, Timothy M. (2002), 'Reasons and Passions', in S. Buss and L. Overton (ed.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge, M.A.: Massachusetts Institute of Technology Press).
Steward, Helen (2012), *A Metaphysics for Freedom* (Oxford: Oxford University Press).
Taylor, Charles (1979), 'What's Wrong with Negative Liberty', in A. Ryan (ed.), *The Idea of Freedom* (Oxford: Oxford University Press).
Taylor, James Stacey (2009), *Practical Autonomy and Bioethics* (Routledge Annals of Bioethics; New York and London: Routledge).
Velleman, J. David (1992), 'What Happens When Someone Acts?', *Mind*, 101, 461–81.
———. (2000), *The Possibility of Practical Reason* (Oxford: Clarendon Press).
——— (2002), 'Identification and Identity', in S. Buss and L. Overton (ed.), *Contours of Agency: Essays on Themes from Harry Frankfurt* (Cambridge, M.A.: Massachusetts Institute of Technology Press).
Watson, Gary (1975), 'Free Agency', *Journal of Philosophy*, 72, 205–20.
——— (1977), 'Skepticism about Weakness of Will', *Philosophical Review*, 86, 316–39.
——— (2004), *Agency and Answerability: Selected Essays* (Oxford: Oxford University Press).
Wertheimer, Alan (1987), *Coercion* (Princeton, N.J.: Princeton University Press).