Philosophical Issues
A Supplement to NOÛS

Philosophical Issues, 22, Action Theory, 2012

RESPONSIBILITY AND AUTONOMY: THE PROBLEM OF MISSION CREEP

John Martin Fischer
University of California, Riverside

I. Introduction

Contemporary philosophers of action have made some breathtaking advances in understanding central features of agency: action, intention, freedom of the will, moral responsibility, and autonomy. Here I wish to pause to reflect on some of these important ideas, especially about acting freely, moral responsibility, and autonomy. I shall suggest that, whereas great progress has been made, we would benefit from keeping firmly in mind some perhaps subtle—but crucial—distinctions. These distinctions can be obscured by the similar language that we use to refer to importantly different phenomena.[1]

II. Frankfurt's Early Account of Acting Freely

In his classic paper, "Freedom of the Will and the Concept of a Person," Harry Frankfurt gives an account of acting freely—the notion of freedom that he thinks is necessary and sufficient for moral responsibility.[2] He famously distinguishes between first-order desires (desires for actions) and second-order desires (desires for first-order desires); among the second-order desires are "second-order volitions"—desires that a certain first-order desire move one all the way to action (and thus constitute the "will"). When one acts in accordance with one's second-order volition, on Frankfurt's view, one acts freely: there is a mesh between one's second-order volition and the first-order desire that moves one all the way to action. That is, the agent has secured a conformity between his second-order volition and his will.

Frankfurt states:

> There is a very close relationship between the capacity for forming second-order volitions and another capacity that is essential to persons—one that has often

> been considered a distinguishing mark of the human condition. It is only because a person has volitions of the second-order that he is capable both of enjoying and of lacking freedom of the will. The concept of a person is not only, then, the concept of a type of entity that has both first-order desires and volitions of the second order. It can also be construed as the concept of a type of entity for whom the freedom of its will may be a problem . . . [3]

It is important to Frankfurt that a person cares about his will, that is, cares about which first-order desire actually moves him effectively to action. More specifically, a person (on Frankfurt's account) identifies with at least some of his first-order desires by forming second-order volitions. He says:

> Now it is having second-order volitions, and not having second-order desires generally, that I regard as essential to being a person. It is logically possible, however unlikely, that there should be an agent with second-order desires but with no volitions of the second order. Such a creature, in my view, would not be a person. I shall use the term 'wanton' to refer to agents who have first-order desires but who are not persons because, whether or not they have desires of the second order, they have no second-order volitions.
>
> The essential characteristic of a wanton is that he does not care about his will.[4]

Now it is very clear that a major aim of Frankfurt's in presenting this account of acting freely is to specify the notion of freedom that is necessary and sufficient for moral responsibility. Perhaps he has other goals, but it seems to me that this is indeed Frankfurt's primary goal in this paper. Indeed, in the final section of the paper, Frankfurt emphasizes that his account of freedom can play an important role in giving an account of moral responsibility. He first argues that his theory of freedom of the will can meet two conditions that various other accounts cannot: it can account for our disinclination to attribute such freedom to members of other species and it can show why freedom of the will is desirable. Frankfurt goes on to say:

> It is generally supposed that, in addition to satisfying the two conditions I have mentioned, a satisfactory theory of the freedom of the will necessarily provides an analysis of one of the conditions of moral responsibility. The most common recent approach to the problem of understanding the freedom of the will has been, indeed, to inquire what is entailed by the assumption that someone is morally responsible for what he has done. In my view, however, the relation between moral responsibility and the freedom of the will has been very widely misunderstood. It is not true that a person is morally responsible for what he has done only if his will was free when he did it. [On Frankfurt's view, a person's will is free only if he could have had a different will—could have willed otherwise.]
>
> . . . This assumption *does* entail that the person did what he did freely . . . It is a mistake, however, to believe that someone acts freely only when he is free to do whatever he wants . . . [5]

Frankfurt makes it clear in the concluding section of the paper that the sort of freedom that involves access to alternative possibilities (freedom to will or do otherwise) is *not* required for moral responsibility.[6] Rather, according to Frankfurt, the freedom-relevant requirement for moral responsibility is acting freely, and—in the view presented in "Freedom of the Will and the Concept of a Person"—it suffices for acting freely that one identifies with the particular first-order desire that moves one to action via the formation of a second-order volition. Here it is in virtue of the second-order volition that one "identifies" with the relevant first-order desire, in the sense of "identification" involved in moral responsibility. We can say that Frankfurt is here answering the question of what must be added to merely acting on desire to get to the kind of freedom linked to moral responsibility; the answer is that one must in a distinctive way "identify" with the desire, and this identification consists in forming an appropriate second-order volition.

Frankfurt concludes the paper with ruminations on the relationship between his account of freedom of the will and the traditional metaphysical worries about the relationship between such freedom and causal determinism:

> My conception of the freedom of the will appears to be neutral with regard to the problem of determinism. It seems conceivable that it should be causally determined that a person is free to want what he wants to want. If this is conceivable, then it might be causally determined that a person enjoys a free will . . .[7]

It is thus indisputable that Frankfurt considers his account of acting freely to be helpful in making progress in traditional debates about causal determinism, freedom of the will, and moral responsibility. Whatever else Frankfurt had in mind in putting forward his "hierarchical" account of acting freely, he certainly sought to make contributions to the traditional metaphysical debates about determinism, free will, and moral responsibility.[8]

III. Two Early Critiques of Frankfurt's Account of Acting Freely

III.1 Thalberg's Critique: Planes of Conation and Stratospheric Yearnings

In his paper, "Hierarchical Analyses of Unfree Action," Irving Thalberg presented various challenges to the leading hierarchical accounts of acting freely on offer at the time.[9] He says:

> Frankfurt attributes great ontological significance to planes of conation within the agent. On his view, 'one essential difference between persons and other creatures' is that persons 'are able to form what I shall call "second-order desires"'. [Here Thalberg refers to Frankfurt, "Freedom of the Will and the

> Concept of a Person", p. 6.] Actually, Frankfurt goes on to distinguish between our second-floor desire that we should merely *have*, or experience, a certain ground-floor desire to act, and our upper-story desire that the ground-level desire 'be the desire that moves [us] effectively to act'. [Frankfurt, "Freedom of the Will and the Concept of a Person", p. 10) The latter kind of stratospheric yearning Frankfurt dubs a 'second-order volition'.[10]

Thalberg presents Frankfurt's discussion of the "unwilling addict" as follows:

> Frankfurt supposes that the addict has a ground-floor disinclination to use narcotics, as well as a craving for them. But from his second-order balcony the fellow 'identifies himself . . . with one . . . of his conflicting desires [the desire not to take his anodyne] . . . makes [it] . . . more truly his own and . . . withdraws himself from the other'; consequently, the 'force moving him to take the drug' must be 'a force other than his own'. [Frankfurt, "Freedom of the Will and the Concept of a Person", p. 13][11]

After presenting a similar discussion by Gerald Dworkin of someone who wants to give up tobacco, Thalberg lays out his worries as follows:

> Why is the drug user's, and the smoker's, craving not really 'his'? Frankfurt and Dworkin seem to be telling us that he does not really want to drug himself, or to smoke, because the real he—his 'true self'—-on its second-order pedestal, abhors these first-order longings. This picture is attractive, but is it cogent?
>
> Both Frankfurt and Dworkin assume that when you ascend to the second level, you discover the real person and what she or he really wants. I shall pack my misgivings into a couple of challenges: Why not go on to third-story or higher desires and volitions? And if that is somehow impossible, why grant that a second-order attitude must always be more genuinely his, more representative of what he genuinely wants, than those you run into at ground level?[12]

Thalberg emphasizes in his paper that it is entirely unclear that the "real self" is the "higher-order" self; in a related (although different) point, he states that it is contentious whether the "real self" is the "rational self". After all, theorists such as Freud have called this into question; Thalberg might have added that much data from social psychology (including the "automaticity" literature and also the "situationism" literature) similarly call into question the notion that the real self is the rational self. As Thalberg puts it, "In commonsense terms, I do not think it is impossible for someone to be a fundamentally irrational person, and to really desire such-and-such while he is in one of his bull-headed moods."[13] He further states that the equation of the real self with the rational self "[begs the question] against Freud's basic hypothesis that other, darker, savage, and nonrational aspects are equally—if not more—important."[14]

III.2 Watson's Critique: The Normative Conception of Agency

In his classic paper, "Free Agency," Watson presented some challenges that are similar to Thalberg's.[15] Watson shares with Thalberg the worry that ascent to the second-level does not help to specify the relevant notion of identification. Thalberg wanted to know why the second-level should be equated with the real self. Similarly, Watson says:

> In a case of conflict, Frankfurt would have us believe that what it is to identify with some desire rather than another is to have a volition concerning the former which is of higher order than any concerning the latter. That the first desire is given a special status over the second is due to its having an n-order volition concerning it, whereas the second desire has at most an (n − 1)-order volition concerning it. But why does one necessarily care about one's higher-order volitions? Since second-order volitions are themselves simply desires, to add them to the context of conflict is just to increase the number of contenders; it is not to give a special place to any of those in contention. The agent may not care which of the second-order desires win out. The same arises at each higher order.[16]

Watson's main point here is that a second-order volition is simply a desire, and, as such, it does not have any special authority or power to "speak for the agent" or indicate where the agent really stands, as it were. Mere desires, according to Watson, even desires at higher orders, are simply "contenders" of the same sort as other desires, and second-order desires do not have any special status, simply in virtue of being higher-order. Watson holds that if we wish to capture the notion of identification at stake, we need to invoke not mere desires, but "values". On Watson's "Platonic" view, some desires stem from evaluations; these desires are not "mere desires", but insofar as they stem from the agent's reasoning about value, they are of a different kind—a kind that can plausibly do the work required by the notion of identification. Watson's key idea here is that we cannot specify the notion of what the agent "really wants" simply by invoking additional desires; in order to reveal what the agent "really wants", in the relevant sense, one has to invoke something akin to Plato's division of the soul. On this view, what an agent really prefers flows in a certain way from the rational part of his soul (speaking metaphorically); that is, what an agent really prefers issues from considerations about what is normatively defensible (in a way that is left unspecified), considering one's life as a whole.[17]

Watson points out that *simply* positing a hierarchy of (mere) desires does not help to specify what an agent really prefers (in the sense at issue). After all, the agent may be a wanton at the second level or at any higher level. Watson also considers Frankfurt's statement, "When a person identifies himself *decisively* with one of his first-order desires, this commitment 'resounds'

throughout the potentially endless array of higher orders . . . "[18] About this Watson says:

> But either this reply is lame or it reveals that the notion of a higher-order volition is not the fundamental one. We wanted to know what prevents wantonness with regard to one's higher-order volitions. What gives these volitions any special relation to 'oneself'? It is unhelpful to answer that one makes a 'decisive commitment,' where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary.[19]

IV. Frankfurt's Reply: Wholeheartedness

IV.1 Decisive Commitment

In work after "Freedom of the Will and the Concept of a Person," Frankfurt has sought to explain what he calls the "resonance effect" and thus assuage at least some of the worries of such philosophers as Watson and Thalberg. In "Identification and Wholeheartedness," he employs an analogy with checking one's answer to a math problem until one is convinced that no further checking will lead to any change in the answer.[20] Frankfurt says:

> The fact that a commitment resounds endlessly *is* simply the fact that the commitment is *decisive*. For a commitment is decisive if and only if it is made without reservation. And making a commitment without reservation means that the person who makes it does so in the belief that no further accurate inquiry would require him to change his mind. It is therefore pointless to pursue the inquiry any further. This is, precisely, the resonance effect.[21]

Frankfurt goes on to apply the analogy with the math problem to the context of practical reasoning and his hierarchical account:

> Now what leads people to form desires of higher orders is similar to what leads them to go over their arithmetic. Someone checks his calculations because he thinks he may have done them wrong. It may be that there is a conflict between the answer he has obtained and a different answer which, for one reason or another, he believes may be correct; or perhaps he has merely a more generalized suspicion, to the effect that he may have made some kind of error. Similarly, a person may be led to reflect upon his own desires either because they conflict with each other, or because a more general lack of confidence moves him to consider whether to be satisfied with his motives as they are.
>
> Both in the case of desires and in the case of arithmetic a person can without arbitrariness terminate a potentially endless sequence of evaluations when he finds that there is no disturbing conflict, either between results already obtained or between a result already obtained and one he might reasonably expect to obtain if the sequence were to continue.[22]

The account of the relevant notion of "identification" that emerges from "Identity and Wholeheartedness" then is roughly as follows. An agent acts freely just in case (a) he acts in accordance with an apparently unopposed volition of order "n" (where n is greater than 1), and (b) he believes that any further reflection (including an ascent to even higher orders) would not issue in any change in his configuration of preferences or reveal a conflict at any level higher than n. When clause (b) is satisfied (along with [a]), the agent has made a "decisive commitment", and, according to Frankfurt, this sort of commitment explains the "resonance effect" in a non-arbitrary way.

IV.2 A Refinement: Satisfaction

In his American Philosophical Association Presidential Address, "The Faintest Passion," Frankfurt essentially expresses some concerns about clause (b) in the above account.[23] Here Frankfurt appears to retreat a bit from the notion that identification involves a "positive element", such as a judgment or belief of the sort specified in clause (b). He worries that the agent might be dissatisfied with the element in question, and that addressing the problem posed by this sort of situation would require positing a problematic infinite number of such deliberate elements (judgments, decision, and so forth).[24] Frankfurt says:

> Being genuinely satisfied is not a matter, then, of choosing to leave things as they are or of making some judgment or decision concerning the desirability of change. It is a matter of simply *having no interest* in making changes. What it requires is that psychic elements of certain kinds *do not occur*. But while the absence of such elements does not require either deliberate action or deliberate restraint, their absence must nonetheless be reflective. In other words, the fact that the person is not moved to change things must derive from his understanding and evaluation of how things are with him. Thus, the essential non-occurrence is neither deliberately contrived nor wantonly unselfconscious. It develops and prevails as an unmanaged consequence of the person's appreciation of his psychic condition.[25]

Frankfurt calls the condition of an agent who has no interest in making changes of the relevant sort "satisfaction." Frankfurt gives the following gloss of this important notion:

> . . . What satisfaction [entails] is an absence of restlessness or resistance. A satisfied person might willingly accept a change in his condition, but he has no active interest in bringing about a change . . . . [A]s a sheer matter of fact, he has no ambition for improvement; he accepts the state of things as it is, without reservation and without any practical interest in how it compares with other possibilities.[26]

This then suggests that Frankfurt accepts the following account of acting freely: An agent acts freely just in case (a) he acts in accordance with an apparently unopposed volition of order "n" (where n is greater than 1), and (b*) he is satisfied with his total psychic economy.[27] Here it is important to Frankfurt that satisfaction not require some "deliberate psychic element" (such as, presumably, "decisive commitment").

Frankfurt applies his account to someone who is trying to quit smoking.[28] As Frankfurt conceptualizes such a situation, the individual has a first-order desire to smoke and a first-order desire to quit smoking. Suppose he also forms a second-order volition to act in accordance with his desire not to smoke; what would indicate that the individual is somehow more closely associated with—identified with—this second-order volition (and the associated first-order desire)? As Frankfurt puts it, echoing Watson, "Considered in itself, after all, his desire to defeat the desire to smoke is just another desire. How can it claim to be constitutive of what he really wants?"[29] Frankfurt's answer to this question is that his approach does not entail that it is simply in virtue of the existence of the second-order volition that the individual identifies with the relevant first-order desire. Nor does such identification require an endless proliferation of positive psychic elements. Rather, as Frankfurt puts it, "Identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied."[30] Perhaps Frankfurt would accept this slight clarification: identification is constituted neatly by an unopposed higher-order endorsing desire in an individual who is satisfied with his psychic economy.[31]

V. Bratman's Critique

Michael Bratman has offered a critique of Frankfurt's notion that a positive mental element—such as a decision—is not required for identification and that the only requirement is something "negative"—such as reflective satisfaction.[32] Bratman imagines that I am reflecting on some higher-order desire and wondering whether to challenge it. As I consider it, my reflections are so far incomplete and have not come to any resolution. Bratman points out that my situation here cannot yet count as identifying with that higher-order desire (or the relevant first-order desire picked out by it), even though I am not (yet) inclined to change anything in light of my understanding of the situation (and thus would seem to meet Frankfurt's conditions for satisfaction). So far we do not have the conditions that would be intuitively sufficient for identification. As Bratman puts it:

> . . . the mere absence of . . . a rejection of my desire is not yet enough for identification. Identification seems to require that I somehow settle the question of the status of my desire. To settle that question, my reflections need to reach

> closure—they need to reach a conclusion. But that seems to mean that my reflections need to reach some sort of decision about whether to challenge that higher-order desire or to 'leave things as they are': the mere absence of motivation to 'change things' seems not to suffice.[33]

Bratman goes on to explain why a mere negative condition, such as Frankfurt's "satisfaction", is not enough to secure "identification":

> . . . one may leave things as they are because of some sort of enervation or exhaustion or depression or the like. If in such a case one has not actually decided to leave things as they are, one has not, I think, identified with how things are with one.[34]

Bratman certainly has a point. How can a psychic element be really one's own—how can one be identified with it (in the relevant sense)—if it is in place because of mere ennui, exhaustion, anomie, or depression? How can this element genuinely "speak for the agent"? In trying to avoid a problematic regress, it seems that Frankfurt has weakened the requirements for identification excessively.

VI. Two Kinds of Identification

VI.1 Some Distinctions

I believe that the literature stemming from Frankfurt's seminal article, "Freedom of the Will and the Concept of a Person," has suffered from a failure to make certain distinctions—or, when these distinctions have been made, to keep them clearly in mind. More carefully, although the contributors to the literature are at least in some cases aware of the distinctions, even they seem sometimes to elide them. But my aim here is not to criticize any of the contributors to this literature; after all, they have made significant advances in understanding central features of our agency. Rather, I wish to make these distinctions explicit and to indicate how a failure to keep them in mind can lead to confusions; even when particular authors do indeed acknowledge the relevant distinctions, it is perhaps easy for a reader to lose track of them. It is also too easy for anyone seeking to understand the relevant phenomena not to attend to the distinctions and thus to make important errors—or at least so I contend.

First, I wish to distinguish between moral responsibility and autonomy. I take it that in his 1971 paper Harry Frankfurt was seeking to give an account of what he believed to be the freedom-relevant component of moral responsibility: acting freely. Of course, when one is morally responsible, one is an appropriate target for certain attitudes, such as indignation, resentment,

gratitude, and respect, and one is also a legitimate target for moral praise and blame. This is, very roughly speaking, what moral responsibility consists in. Further, I take it genuine weakness of the will is possible; that is, I assume that one can act freely against the dictates of one's "true self" or "real self" (I'll say more about the true or real self below). When one exhibits weakness of the will, one acts freely and is morally responsible; indeed, one is criticizably irrational for the weak-willed behavior. Thus, an agent's being morally responsible for a bit of behavior is perfectly consistent with an agent's falling short of certain norms of rationality (as well as morality) in so behaving; it is obvious that our conceptualization of moral responsibility must allow for moral responsibility for morally wrong actions.

I take it that the more robust notion of autonomy is inconsistent with weakness of the will (and thus with failing to act in accordance with the relevant norm of rationality). Autonomy entails moral responsibility, but it also requires that an additional condition be met; autonomy is "self-governance", and thus it requires (in some difficult-to-specify sense) governance by the self. However precisely one characterizes the relevant self-governance, I presume that it rules out weakness of the will. Thus, I will use the term "autonomy" to pick out a notion that is more demanding than mere moral responsibility, which is compatible with weakness of the will. More specifically, it requires a kind of governance or direction by the "self", where the self is here understood as the "real self" (for practical purposes) or perhaps the agent's "practical identity". The pertinent notion of "real self" or "agential identity" is not meant to bring with it heavy ontological baggage; rather, it is intended to pick out the "designated" part of an agent's psychic economy that somehow most truly or deeply indicates where the individual stands (for practical purposes).

I contend that an agent can be fully morally responsible for an action that is not autonomous. Various philosophers employ different terms for features of autonomous agency. For example, David Velleman employs the term "agency [or action] *par excellence*."[35] And Michael Bratman talks about "agential authority". When a higher-order desire has "agential authority", on Bratman's view, it "has authority to speak for the agent, to constitute where the agent stands."[36] For Bratman, when an agent acts autonomously, he acts in accordance with a higher-order desire that grants agential authority—one that shows or constitutes where the agent (really) stands. Bratman is quite clear that there is a distinction between autonomy and moral responsibility, and that his focus is on autonomy:

> . . . I understand self-governance of action to be a distinctive form of self-direction or self-determination (I do not distinguish these last two) of action. Autonomy—that is, personal autonomy—is self-direction that is, in particular, self-governance. Or anyway, that is the phenomenon that is my concern here. [Bratman here refers to his, "Autonomy and Hierarchy," in Ellen Frankel Paul,

Fred D. Miller, jr., and Jeffrey Paul, eds., *Autonomy* (New York: Cambridge University Press, 2003); 156–176, esp. pp. 156–7 and 168.] Autonomy is related in complex ways to moral responsibility and accountability, but I do not consider these further issues here.[37]

For simplicity's sake, I shall use "autonomy" to cover the various notions that require action in accordance with the "true" or "real" self—the self that somehow stands for the agent or is the perspective or location or standpoint of the agent, for the purposes of practical reasoning.

Whereas most philosophers are aware of the distinction between moral responsibility and autonomy (conceived as above), perhaps not so many see a corresponding distinction between two kinds of "internality" or "identification". I claim that there are indeed two kinds of "internality" or "identification", corresponding to the two distinct notions of moral responsibility and autonomy. I'll put the distinction in terms of identification.

Call the notion of identification relevant to moral responsibility, "responsibility-identification", and the notion of identification relevant to autonomy, "autonomy-identification". I contend that these two kinds of identification are distinct. Further, I contend that whereas autonomy-identification is naturally specified in terms of the "true" or "real" self—the elements or structures of the self that somehow show where the agent really stands, as above—the issue of what constitutes responsibility-identification is an open question. At least, it is not obvious that responsibility-identification requires autonomy-identification, conceptualized in terms of a link with the agent's "true self" (as understood above). After all, one can be morally responsible for weak-willed behavior and also unreflective behavior, and even behavior that is significantly out of character.

VI.2 Applications

I claim that one's analysis of the target phenomena can be vitiated significantly by failing to keep separate these (arguably) separate notions of identification (as well as the paired notions of moral responsibility and autonomy). Begin with Thalberg's critique. I grant that it is fair to ask whether the relevant higher-order preference corresponds to some sort of "real" self, understood either in a robust metaphysical sense or more modestly as the true self with respect to practical reasoning. Similarly, it is contentious whether "reason" or "rationality" corresponds to the real self (understood robustly or modestly).

Frankfurt himself is at pains to point out that the relevant higher-order preference need not be based on "reason" in any sense, in order for it to play the requisite role in his account; the higher-order preference in question can be formed on the basis of no reflection at all, in Frankfurt's view. So it is clear

that Frankfurt would not wish to assimilate the real self to the rational self. Further, no proponent of the hierarchical model of acting freely need say that the relevant higher-order preference corresponds to the real self (understood as where the agent really stands, with respect to practical reasoning). Rather, the proponent of such an account of acting freely may say that the relevant higher-order preference is part of an analysis of responsibility-identification, *not* autonomy-identification. If this is correct, then Thalberg's critique of Frankfurt's hierarchical approach—and all hierarchical accounts of acting freely—misses the mark. That is, it misses the mark insofar as the target phenomena are responsibility-identification and moral responsibility, rather than autonomy-identification and autonomy. At most the critique shows that the hierarchical accounts fall short as analyses of autonomy-identification and autonomy.

Consider also Watson's critique. Although there are various elements to it, a fundamental point made by Watson is that higher-order preferences are "mere desires," and, as such, are not suited to speak for the self any more than other desires, such as first-order preferences. On Watson's Platonic view of the "soul" or self, there are fundamentally different "parts" or sources of motivation, and the rational part or source of motivation has hegemony in specifying who we are as agents—where we stand, as it were. Watson contends that mere motivations—desires that do not stem in the appropriate sense from a certain sort of rational reflection—cannot, by their very nature, constitute what we really want.

Watson's point might be correct with respect to the true self, understood (roughly) as where the agent really stands with respect to practical reasoning. On this sort of view, a mere desire cannot "speak for the agent" in the relevant sense. Thus, if Watson is indeed correct here, this would be relevant to autonomy-identification and autonomy. But, as with my analysis of Thalberg's critique above, it is completely unclear whether Watson's point applies to responsibility-identification and moral responsibility. Minimally, I would contend that it does not *follow* from its applicability to the autonomy phenomena that it also applies to the responsibility phenomena.[38]

Note further that it is notoriously difficult for Watson's sort of account of acting freely to accommodate the phenomenon of weakness of the will. As I stated above, I take it that there are genuine cases of weakness of the will—cases in which an agent acts freely against what he takes to be the best thing to do (all-things-considered). It is difficult for Watson's approach to allow for weakness of the will, since, on his approach, one acts freely just in case one acts (in a sense left somewhat vague) in accordance with reason. In any case, Watson is aware of these difficulties, and perhaps there are ways of developing his sort of "normative" account of acting freely that fit with the possibility of weakness of the will.[39] It suffices here to observe that, given the phenomenon of weakness of the will, Watson's approach

is more plausible as an account of autonomy-identification and autonomy than responsibility-identification and responsibility. A weak-willed agent is perhaps not autonomous; he does not exhibit "agency *par excellence*" and his actions do not point to where he really stands. But he nevertheless acts freely in the sense relevant to moral responsibility.

Consider this interesting passage from Watson:

> One's evaluational system may be said to constitute one's standpoint, the point of view from which one judges the world. The important feature of one's evaluational system is that one cannot coherently dissociate oneself from it *in its entirety*. For to dissociate oneself from the ends and principles that constitute one's evaluational system is to disclaim or repudiate them, and any ends and principles so disclaimed (self-deception aside) cease to be constitutive of one's valuational system. One can dissociate oneself from one set of ends and principles only from the standpoint of another such set that one does not disclaim. In short, one cannot dissociate oneself from all normative judgments without forfeiting all standpoints and therewith one's identity as an agent.[40]

I can here stipulate, for the sake of the discussion, Watson's view that one's evaluational system constitutes one's "identity as an agent" and also Watson's further point that one cannot dissociate oneself from all evaluational judgments and still be an agent. Still it does not follow that, in any particular action, an agent (so understood) must act in accordance with his evaluative judgment as to what is best to do. He could perform a weak-willed action. Also, he could simply act spontaneously or from a whim, not informed or guided by his evaluative system at all. As I would interpret the situation, the agent can indeed responsibility-identify with such actions; he can thus be morally responsible for his weak-willed, unreflective, or capricious actions. The fact that, *qua* agent, he must be conceptualized as *having* at least some evaluative standpoint does not entail that in a particular case his action *stems from* this sort of standpoint. One can agree with Watson about the impossibility of dissociating oneself entirely with the normativity or our evaluative or practical commitments while maintaining the distinction between the responsibility phenomena and the autonomy phenomena.

Think of autonomy as self-governance in the sense of direction by the "true" self—the "inner citadel"[41] or "designated part of the self", for the purposes of practical reasoning. Watson and others have specified Reason (in various specific ways) as the inner citadel or designated part of the self. So autonomy is governance by Reason. Again, I stipulate this point for the sake of argument here. But, again, I insist that it does not follow that responsibility-identification and moral responsibility should be conceptualized as governance by Reason; insofar as the responsibility phenomena are distinguished from the autonomy phenomena, this remains an open question.[42]

We might think of autonomy as self-governance in the sense of direction by the "true self" but not conceive of the true self in terms of rationality or normativity. For example, Bratman thinks of the true self as specified (very roughly speaking) by the agent's long-term plans.[43] Bratman argues that his view—the Planning Theory—is importantly different from Watson's Reason View. But, as above, because moral responsibility must be distinguished from autonomy, it would not follow that responsibility-identification and moral responsibility should be interpreted as governance in accordance with long-term planning structures.[44]

Recall, also, Bratman's critique of Frankfurt's notion of satisfaction. Bratman points out that we could be satisfied with our psychic condition, in Frankfurt's sense, simply because we are tired (or even exhausted) or depressed. It is difficult for Bratman to see how being satisfied in these ways could help to specify "agential authority" or what truly speaks for the agent. One can agree with Bratman, insofar as the target phenomena are autonomy-identification and autonomy; and recall that Bratman himself is clear about the distinction between autonomy and responsibility and takes his target to be autonomy.

Here I wish simply to point out that Bratman's critiques of Frankfurt would *not* apply to Frankfurt's account insofar as the target is the responsibility phenomena, rather than the autonomy phenomena. Perhaps it is clear that if the agent is satisfied with the relevant higher-order psychological structures simply because he is (say) tired, bored, or depressed, we do not yet know where the agent really stands, in the sense relevant to autonomy. But it certainly does not follow that we do not have an adequate account of responsibility-identification, as it constitutes the freedom-relevant condition on moral responsibility. From the mere fact that an agent is to some (even a significant) degree depressed, and the fact that this depression plays a role in his lack of interest in changing his preference structure, it does *not* follow that the agent cannot be morally responsible for the relevant action. Similarly the mere facts of exhaustion or boredom or anomie or alienation do not in themselves remove moral responsibility. One might ask why exactly the agent is alienated. Why is he tired? Why doesn't he care? An agent can certainly be accountable for being tired, inattentive, or for not caring enough, and thus it remains an open question that satisfaction that emerges from such conditions might be part of an adequate account of moral responsibility. It would be a mistake, although not a mistake made by Bratman himself, to extrapolate from a critique of Frankfurt, interpreted as giving an account of the autonomy phenomena, to a critique of Frankfurt, interpreted as giving an account of the responsibility phenomena. And, as I pointed out above, Frankfurt takes himself—at the very least—to be making a contribution to traditional debates about causal determinism, freedom of the will, and moral responsibility; whether he also wishes to offer an analysis of autonomy is unclear.

VII. Conclusion: Mission Creep

In his great 1971 paper, "Freedom of the Will and the Concept of a Person," Harry Frankfurt set out to give a more refined account of acting freely than the standard compatibilist accounts on offer. Frankfurt wished (among other things) to give a compatibilist-friendly account of acting freely that specifies what has to be added to mere action from a first-order desire to get to "acting freely" in the sense implicated in moral responsibility. On this approach, what is required is the presence of certain higher-order psychological structures.

Frankfurt's project here was very similar to that of Gary Watson in his remarkable 1975 paper, "Free Agency". Watson says:

> Now, though compatibilists from Hobbes to J.J.C. Smart have given the relevant moral and psychological concepts an exceedingly crude treatment, this crudity is not inherent in compatibilism . . .
>
> In the subsequent pages, I want to develop a distinction between wanting and valuing which will enable the familiar view of freedom [as the ability to get what one wants] to make sense of the notion of an unfree action. The contention will be that, in the case of actions that are unfree, the agent is unable to get what he most wants *or values*, and this inability is due to his own 'motivational system.' . . .
>
> I do not conceive my remarks to be a defense of compatibilism. This point of view may be unacceptable for various reasons, some of which call into question the coherence of the concept of responsibility. But these reasons do not include the fact that compatibilism relies upon the conception of freedom in terms of the ability to get what one wants . . .[45]

So in their classic early papers Frankfurt and Watson seemed to be on the same page, as it were: they were seeking to provide more sophisticated accounts of freedom than the "familiar" idea that one is free to the extent that one has the ability to get what one wants. Although neither philosopher explicitly endorsed compatibilism, they both conceived of their projects as part of defending compatibilism against certain objections.

Various philosophers also talk about "agency *par excellence*", "autonomy", and action governed or directed by the agent's "real" or "true" self, or perhaps by psychological elements or structures that somehow indicate or constitute the agent's true standpoint, where the agent really stands, or have 'agential authority". Here the philosophers in question obviously have in mind something more robust than what is required for moral responsibility. Above I said that for simplicity's sake, I assume that "autonomy" (and autonomous action) requires direction by the agent's "real" or "true" self, whereas moral responsibility need not require direction by

such "designated" elements. I claimed that there is a distinctive and separate kind of identification associated with autonomy and moral responsibility: autonomy-identification and responsibility-identification.

Another way to regiment the pertinent range of phenomena would have it that there is a "true self" or "designated element or structure" for each phenomenon—autonomy and moral responsibility. On this view, autonomy-identification involves the true or designated self, with respect to autonomy. In contrast, responsibility-identification is specified in terms of the true or designated self, with respect to moral responsibility. Further, I contend that these true or designated selves need not be the same; it would be expected that the autonomy-self (conceived of as a designated element or structure) is more robust or has more stringent requirements than the responsibility-self (also conceived of designated elements or structures). Responsibility-identification here presupposes that there is a notion of an action "speaking for me" in a way that licenses "reactive attitudes," such as indignation, resentment, gratitude, and respect, which is compatible with weakness of the will; this obviously contrasts with the autonomy sense of "speaking for me" which is taken to be incompatible with weakness of the will.

So, one way of conceptualizing the phenomena posits two kinds of identification, in which only autonomy-identification involves the true or designated self; the other way envisages two kinds of identification, in which two kinds of true or designated self play the crucial role. In any case, once one separates autonomy from moral responsibility—and the attendant notions of identification—one can see that objections to a particular account of identification, as it relates to autonomy, are not *eo ipso* objections to that account of identification, as it relates to moral responsibility. Because of the close similarity of the relevant notions of identification, and also because of the similarity of the language we use to describe the relevant phenomena, it is perhaps easy to slide from one notion to the other without noticing.[46]

At the risk of etiolating whatever scholarly reputation I might have, I would like to point to the beginning of the "Wikipedia" entry on "mission creep":

> Mission creep is the expansion of a project or mission beyond its original goals, often after initial successes. The term often implies a certain disapproval of newly adopted goals by the user of the term. Mission creep is usually considered undesirable due to the dangerous path of each success breeding more ambitious attempts, only stopping when a final, often catastrophic, failure occurs. The term was originally applied exclusively to military operations, but has recently been applied to many different fields. The phrase first appeared in articles concerning the UN Peacekeeping mission during the Somali Civil War in the Washington Post on April 15, 1993 and in the New York Times on October 10, 1993.[47]

I do think that there has been a certain expansion of the project of Frankfurt and Watson beyond its original goals. I wish in this paper simply to draw

attention to the distinct goals of giving accounts of the different—but easily conflated—target phenomena: responsibility and autonomy. I certainly do not think it is dangerous or potentially catastrophic to seek to give an illuminating account of the more robust notion of autonomy. But I do think that it can be philosophically dangerous—even if not entirely catastrophic—to fail to distinguish this project from the related but distinct project of giving an account of moral responsibility and thus to slide (perhaps imperceptibly) from one philosophical goal to another.[48]

Notes

1. I have made a preliminary effort to sketch the ideas I shall develop more fully in the current paper in: John Martin Fischer, "Responsibility and Autonomy," in T. O'Connor and C. Sandis (eds), A Companion to the Philosophy of Action (Oxford: Wiley/Blackwell, 2010), pp. 309–16.
2. Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68 (1971), pp. 5–20; reprinted in John Martin Fischer, ed., *Moral Responsibility* (Ithaca: Cornell University Press, 1986), pp. 65–80. All references here will be to the reprinted version. In a previous paper, Frankfurt had argued that freedom to do otherwise is not necessary for moral responsibility: "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66 (1969), pp. 828–39; reprinted in Fischer, ed., *Moral Responsibility*, pp. 143–52.
3. Frankfurt, "Freedom of the Will and the Concept of a Person," p. 74.
4. Frankfurt, "Freedom of the Will and the Concept of a Person," pp. 70–1.
5. Frankfurt, "Freedom of the Will and the Concept of a Person," pp. 78–9.
6. Of course, he had also argued for this claim in Frankfurt, "Alternate Possibilities and Moral Responsibility."
7. Frankfurt, "Freedom of the Will and the Concept of a Person," p. 80.
8. In "Reply to John Martin Fischer," Frankfurt further elaborated and explained his views on the traditional problem of the relationship between causal determinism and moral responsibility: Harry Frankfurt, "Reply to John Martin Fischer ["Frankfurt-Style Compatibilism"], in Sarah Buss and Lee Overton, eds., *Contours of Agency* (Cambridge, Ma.: MIT Press, 2002), pp. 27–31. Here he explains that his goal has been to address certain objections to compatibilism, although he remains officially neutral about the compatibility of causal determinism and moral responsibility insofar as he is not sufficiently confident that causal determination is consistent with being active (rather than passive): p. 29.
9. Irving Thalberg, "Hierarchical Analyses of Unfree Action," *Canadian Journal of Philosophy* 8 (1978), pp. 211–226.
10. Thalberg, "Hierarchical Analyses of Unfree Action," p. 214.
11. Thalberg, "Hierarchical Analyses of Unfree Action," p. 219.
12. Thalberg, "Hierarchical Analyses of Unfree Action," pp. 219–20. Dworkin's discussion is in: Gerald Dworkin, "Acting Freely," *Nous* 4 (1970), pp. 367–83.
13. Thalberg, "Hierarchical Analyses of Unfree Action," p. 223.
14. Thalberg, "Hierarchical Analyses of Unfree Action," p. 224.

15. Gary Watson, "Free Agency," *Journal of Philosophy* 72 (1975), pp. 205–20; reprinted in Fischer, ed., *Moral Responsibility*, pp. 96.
16. Watson, "Free Agency," pp. 93–4.
17. Watson says, "We might say that an agent's values consist in those principles and ends which he—in a cool and non-self-deceptive moment—articulates as definitive of the good, fulfilling, and defensible life." ("Free Agency," p. 91).
18. The quotation is from Frankfurt, "Freedom of the Will and the Concept of a Person," p. 76.
19. Watson, p. 94.
20. Harry Frankfurt, "Identification and Wholeheartedness," Ferdinand D. Schoeman, ed., *Responsibility, Character and the Emotions* (Cambridge: Cambridge University Press, 1987), pp. 27–45; reprinted in John Martini Fischer and Mark Ravizza, eds., *Perspectives on Moral Responsibility* (Ithaca: Cornell University Press, 1993), pp. 170–87. (All references will be to the reprinted paper.)
21. Frankfurt, "Identification and Wholeheartedness," p. 180.
22. Frankfurt, "Identification and Wholeheartedness," p. 180.
23. Harry G. Frankfurt, "The Faintest Passion," *Proceedings of the American Philosophical Association* 66 (1992), pp. 5–16; reprinted in John Martin Fischer, ed., *Free Will: Critical Concepts in Philosophy* Volume IV (London: Routledge, 2005), pp. 54–67. (All references will be to the reprinted paper.)
24. Frankfurt, "The Faintest Passion," p. 63.
25. Frankfurt, "The Faintest Passion," p. 63.
26. Frankfurt, "The Faintest Passion," p. 62.
27. Thus, satisfaction (for Frankfurt) can involve the presence of an opposing lower-level desire that the agent would rather not have.
28. Frankfurt, "The Faintest Passion," p. 64.
29. Frankfurt, "The Faintest Passion," p. 64.
30. Frankfurt, "The Faintest Passion," p. 64.
31. By "endorsing" here I do not mean to imply any kind of positive evaluation; rather, to endorse, in the relevant sense here, is simply to "throw one's weight behind" something (to employ what might, admittedly, be an unhelpful metaphor). A desire may be said to endorse insofar as it plays an appropriate functional role in an agent who endorses. Here I am indebted to helpful comments by Justin Coates.
    Also, Benjamin Mitchel-Yellin has pointed out to me (personal correspondence) that it is unclear that Frankfurt requires that the relevant higher-order element be "unopposed". Nothing in my discussion in this paper depends on resolving this exegetical question.
32. Michael E. Bratman, "Identification, Decision, and Treating as a Reason," *Philosophical Topics* 24 (1996), pp. 1–18.
33. Bratman, "Identification, Decision, and Treating as a Reason," p. 7.
34. Bratman, "Identification, Decision, and Treating as a Reason," p. 7. He reiterates this critique of Frankfurt in: Michael E. Bratman, "Planning Agency, Autonomous Agency," in James Stacey Taylor, ed., *Personal Autonomy* (Cambridge: Cambridge University Press, 2004), pp. 33–57; reprinted in Fischer, ed., *Ciritical Concepts: Free Will*, Vol. IV, pp. 68–89. (All references will be to the

reprinted paper.) In the aforementioned papers (and others) Bratman develops his own account of "identification" which, he contends, avoids the problem he has specified for Frankfurt's approach.

35. See, for example, J. David Velleman, "What Happens When Someone Acts?" *Mind* 101 (1992), pp. 461–81; reprinted in Fischer and Ravizza, eds., *Perspectives on Moral Responsibility* (Ithaca: Cornell University Press, 1993), pp. 188–210.
36. Bratman, "Planning Agency, Autonomous Agency," p. 72. For the notion of "agential authority," see also Michael E. Bratman, "Two Problems About Human Agency," *Proceedings of the Aristotelian Society* 101 (2001), pp. 309–26.
37. Bratman, "Planning Agency, Autonomous Agency," p. 85, fn. 1.
38. Since Frankfurt takes Watson's critique so seriously, it is interesting to consider whether Frankfurt himself sees the distinction between moral responsibility and autonomy as clearly as he should (or is sufficiently attentive to it). Also, it is unclear exactly how Watson conceives of the relationship between moral responsibility and autonomy (or self-governance). In "Free Action and Free Will" (*Mind* 96 [1987], pp. 145–72; reprinted in Watson, ed., *Agency and Answerability*, 161–98), he claims that an account of free will has two components: an account of autonomy (or self-direction) and an account of alternative possibilities. He claims that compatibilists and incompatibilists start from different places and so place different emphases on each component. This suggests that Watson thinks that autonomy is importantly connected with moral responsibility (perhaps in complex ways). Here I am indebted to comments and also unpublished work by Benjamin Mitchell-Yellin.
39. Gary Watson, "Skepticism About Weakness of the Will," *Philosophical Review* 86 (1977), pp. 316–39; reprinted in Gary Watson, ed., *Agency and Answerability* (Oxford: Clarendon Press, 2004), pp. 33–58; and "Free Action and Free Will".
40. Watson, "Free Agency", pp. 91–2.
41. John Christman begins his helpful introduction to his collection, *The Inner Citadel*, with this quotation from Isaiah Berlin:

> I wish my life and decisions to depend on myself, not on external forces of whatever kind. I wish to be the instrument of my own, not other[s'']acts of will. I wish to be a subject, not an object . . . I wish to be somebody, not nobody.
>
> It is as if I had performed a strategic retreat into an inner citadel—my reason, my soul, my 'noumenal' self—which, do what they may, neither external blind force, nor human malice, can touch. I have withdrawn into myself; there, and there alone, I am secure.

The quotation is from "Two Concepts of Liberty," in Isaiah Berlin, *Four Essays on Liberty* (Oxford: Oxford University Press, 1969), pp. 118–192, esp. pp. 131 and 135. Christman quotes these passages on p. 3 of *The Innter Citadel* (New York, Oxford University Press, 1989).

In a statement that is related to the first part of the passage from Berlin, the comedian, Lily Tomlin once said, "When I was young I just wanted to be somebody. Now I wish I had been more specific."

42. In "Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will," Eleonore Stump supplements Frankfurt's account of acting freely by requiring that the crucial second-order volition be based on the agent's representation of the relevant course of action as good. (*Journal of Philosophy* 85 (1988), pp. 395–420; reprinted in Fischer and Ravizza, eds., *Perspectives on Moral Responsibility*, pp. 211–234. All references will be to the reprinted paper.) Stump here combines Frankfurt's hierarchical model with an element similar to Watson's notion of Value. She argues that the addition of such an element renders Frankfurt better able to avoid criticisms to the effect that his approach does not adequately capture the phenomenon of governance by the true self. (pp. 223–26) But as in my discussion of Watson in the text, even if Stump is correct about the true self and thus autonomy, it would not follow that Frankfurt's account, interpreted as an account of moral responsibility, is inadequate.
43. Bratman, "Planning Agency, Autonomous Agency." Also, see the introductory essay to *Structures of Agency*.
44. I reiterate that Bratman himself does not conflate responsibility and autonomy.
45. Watson, "Free Agency, pp. 82–3.
46. Pamela Hieronymi makes a similar point in "Making a Difference, " *Social Theory and Practice* 37 (2011), pp. 81–94. It is indeed striking how parallel and significantly isomorphic literatures have developed in contemporary philosophy pertaining to autonomy and moral responsibility. In many cases the literatures use identical language. So, just for example, an important early paper on autonomy by Gerald Dworkin is entitled, "Acting Freely" (*Nous* 4 [1970], pp. 367–83). The fact that the parallel literatures on autonomy and responsibility use such similar language can make it difficult to keep the distinctions between the phenomena firmly in mind. And it is not just the language that is similar; the parallel literatures are strongly isomorphic with respect to such issues as whether agency should be conceptualized in a hierarchical fashion, how to accommodate the possibility of problematic manipulative induction of mental precursors to action, and whether agency (of the relevant sort) is essentially a historical notion (just to mention three salient examples). For some classic and important works in the autonomy literature, see: Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge: Cambridge University Press, 1988); Marina Oshana, *Personal Autonomy in Society* (Aldershot, U.K.: Ashgate Press, 2006); and James Stacey Taylor, *Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy* (New York: Cambridge University Press, 2005).
47. http://en.wikipedia.org/wiki/Mission_creepm, accessed June 1, 2011.
48. I am extremely grateful to comments by Agnieszka Jaworska, Benjamin Mitchell-Yellin, Christopher Franklin, and Justin Coates. Also, I have benefited from reading (as yet) unpublished work: Justin Coates, "Nietzsche's Theory of Freedom of the Will in *Beyond Good and Evil*," (manuscript, Department of Philosophy, University of California, Riverside); and Benjamin Mitchell-Yellin, "Self-Governance, Moral Responsibility, and Weakness of Will: In Defense of the Platonic Model" (Department of Philosophy, University of California, Riverside).