forthcoming in *Mental Agency*, Lucy O'Brien, ed. (Oxford: Oxford University Press).

Two Kinds of Agency

Pamela Hieronymi
hieronymi@ucla.edu
April 9, 2008

I will argue that making a certain assumption allows us to conceptualize more clearly our agency over our minds. The assumption is this: certain attitudes (most uncontroversially, belief and intention) embody their subject's answer to some question or set of questions. I will first explain the assumption and then show that, given the assumption, we should expect to exercise agency over this class of attitudes in (at least) two distinct ways: by answering for ourselves the question they embody and by acting upon them in ways designed to affect them according to our purposes—in roughly the way we exercise agency over most ordinary objects.

The two forms of agency are rarely distinguished, because the first does not display the most familiar and prominent features of agency, while the second might involve an exercise of the first, at two distinct points. Nonetheless, many complex exercises of agency over our minds are easily seen—I think best seen—as composed of these two, more simple, forms. My hope is that decomposing the complex exercises of agency into these two forms might bring some clarity to the difficult topic of mental agency.

THE ASSUMPTION

I begin by explaining the assumption. Note that, having settled for oneself some question, one is then in a certain kind of state of mind—namely, a state of mind of having settled that question. For the settling of certain sorts of questions, we give a name to such states. For example, having settled for oneself (positively) the question of whether to ϕ

(where ϕ stands for some ordinary action, such as make some lunch or dust the furniture), one therein intends to ϕ.

Note, too, that (for persons, or rational subjects) insofar as one intends to ϕ, one is vulnerable to certain sorts of criticisms and open to certain sorts of questions—in particular, one is open to questions and criticisms that would be satisfied by reasons that (one takes to) bear positively on whether to ϕ.[1] I will capture this vulnerability with the notion of *commitment*, saying that insofar as a person intends to ϕ, that person is committed to ϕ-ing. In fact, given that the reasons that would satisfy the questions and criticisms to which one is vulnerable are just those that (one takes to) bear positively on whether to ϕ, it seems that, insofar as one intends to ϕ, one is committed to a positive answer to the question of whether to ϕ.

Thus, if one has settled for oneself positively the question of whether to ϕ, one intends to ϕ, and one intends to ϕ just in case one is committed to a positive answer to the question of whether to ϕ. I will capture this complex conjunction of conditionals by saying that an intention to ϕ *embodies one's answer to the question* of whether to ϕ.

It seems these same claims hold of belief. If one settles for oneself positively the question of whether *p* (where *p* stands for a proposition, such as "The butler did it" or "All cats are sweet-tempered, deep down"), then one believes *p*. Likewise, insofar as one believes *p*, one is committed to a positive answer to the question of whether *p*, i.e., one is

[1] I insert the parenthetical "(one takes to)" because certain of the questions and criticisms (such as Anscombe's famous why-question) would be satisfied simply by whatever one took to settle the question, while other questions and criticisms (such as certain kinds of moral criticisms) would be satisfied only by reasons that in fact settle the question, while still others (such as certain concerns about justification) would be satisfied by reasons that would settle the question, given your (actual or idealized) epistemic situation. While this complexity is important, for the matter at hand what is crucial is that the questions and criticisms would all be satisfied by considerations that either do bear, would (given certain assumptions) bear, or were taken to bear on a certain question, namely, whether to ϕ.

vulnerable to a range of questions and criticisms that would be satisfied by reasons that (one takes to) bear positively on whether *p*. So we can say that a belief that *p* embodies a positive answer to the question of whether *p*.

Far more controversially, I think the same sort of claims can be made about certain emotions—that, e.g., one's resentment of *S* for ϕ-ing embodies one's answer to some range of questions about *S*'s ϕ-ing. I will not defend this more controversial claim, here. Though I think it illuminating—it can shed light both on the nature of certain of our emotions and on the nature or status of the claims I have made about belief and intention—the more controversial claim is not needed, for the main point at hand.

TWO KINDS OF AGENCY

If an attitude embodies our answer to a question or set of questions, then it seems we will form or revise such an attitude in forming or revising our answers to the relevant question(s). As noted, if you become convinced that *p*, and so settle for yourself the question of whether *p*, you therein, *ipso facto*, believe *p*. Likewise, if you settle (positively) the question of whether to ϕ, you therein, *ipso facto*, intend to ϕ. Moreover, if you change your mind about whether to ϕ, or about whether *p*, in such a way that you are no longer committed to ϕ-ing or to the truth of *p*, then you no longer intend to ϕ or believe that *p*. We might say that we control these aspects of our minds because, as we change our mind, our mind changes—as we form or revise our take on things, we form or revise our attitudes. I call this exercising *evaluative control* over the attitude.[2]

[2] This is what Richard Moran sometimes calls "deliberative" or "rational" control. See Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press, 2001), especially 113–20. I do not follow him in using that label, since it seems to me to suggest that this kind of agency requires deliberation or reasons.

Though this is, I think, the ordinary and most fundamental way of controlling these attitudes, it is far from an ordinary notion of control or agency. In fact, there are a number of important questions one might raise about it and a number of important objections to calling it a form of control. I will examine some of these objections after considering another form of agency we also exercise with respect to these attitudes.

Note that these attitudes, which (I claim) embody a person's answer to a question or range of questions, also interact in more-or-less predictable ways with their environment. Our attitudes share this feature with ordinary objects, like chairs, coffee cups, and computers. Insofar as we can think about these attitudes and understand their interaction with their environment, we can control them in the same way we control anything that we can think about that interacts in more-or-less predictable ways with its environment: we can take actions designed to affect them according to our purposes. Our ability to thus control our attitudes is limited only by our cleverness, strength, luck, and industry, i.e., by the same features that limit our control over any object. Thus these attitudes can be objects of a far more familiar sort of control, which I call *managerial* or *manipulative control*.

While it might seem surprising that we can exercise the same form of control over our attitudes that we exercise over more ordinary objects, it should not. Consider the relative ease with which we exercise this familiar form of control over the attitudes of others. If you want to bring it about that someone else believes p or intends to ϕ, you will not, generally, be at a loss as to how to proceed. Of course, in certain cases, for certain values of p or ϕ, the task may be too difficult to achieve. But for a great many values, it will be quite doable—you must simply bring it about that the person settles positively the

question of whether *p* or whether to ϕ, and there is a familiar range of ways to accomplish this. It should not be surprising, then, that we can exercise the same sort of control over our own attitudes—that we can take steps designed to bring it about that we believe *p* or intend to ϕ. In order to succeed, we have to bring it about that we have answered positively the question of whether *p* or whether to ϕ—we have to bring it about that we are committed to *p* as true or to ϕ-ing. In certain cases, for certain values of *p* or ϕ, this will be quite difficult. For others, it will be relatively obvious what to do.

There is, of course, a kind of difficulty in one's own case that one does not encounter in managing or manipulating the attitudes of others: In order to bring it about that another person believes or intends, you might provide that person with considerations you predict that person will find compelling, which considerations you do not, yourself, find compelling. But in order to bring it about that you, yourself, believe or intend, and to do so by providing yourself with reasons, you must provide yourself with reasons that you predict you will, yourself, find compelling. But, of course, if you thought there were available compelling reasons, it would be likely that you *already* believe or intend. So the opportunities for managing one's own attitudes by providing oneself with compelling reasons will be more restricted than the opportunities to do so to another. Still, they can arise: If, unable to sleep, you want to believe that your children arrived home safely through the storm, you might call them and so provide yourself with convincing evidence that they have arrived.[3] If you want to be sure that, tomorrow, you will still intend to avoid desert, you might act, today, to create extra incentives: you might make bets with your friends. Moreover, providing reasons for yourself is not the only way in which you

[3] I owe this helpful example to Thomas P. Kelly.

might manage or manipulate a belief or intention. You might undergo hypnosis, or induce amnesia, or convince yourself that an alternative interpretation of your situation is equally justified, and so successfully change your attitudes.

Thus it seems we can manage or manipulate our own attitudes in roughly the way we can manage or manipulate ordinary objects: by taking actions designed to affect them according to our purposes.

EVALUATIVE CONTROL AND OBJECTIONS THERETO

I return, now, to evaluative control. I claimed that certain attitudes embody one's answer to a question or set of questions, and that, therefore, one can exercise control or agency over such attitudes by coming to or revising one's answers to the relevant question(s). I acknowledged that it is a far-from-ordinary notion of control. I will here briefly consider a few objections to it, hoping thereby to display its operation more clearly.[4]

[4] There are two possible ways of elaborating upon this view about our agency over our attitudes. On the first, we would distinguish between *settling* a question and *being committed to* an answer to that question. Settling the question, one might think, it is an activity that one may or may not engage in; being committed to an answer to a question is not an activity, but rather some sort of "normative status"—one is committed just in case one is open to characteristic certain sorts of questions and criticisms (again, questions and criticisms that would be satisfied by reasons (one takes to) support a certain answer to a question). If one has this "normative status," it seems we can say that one is committed to an answer to this question. Thus, on this interpretation of evaluative control, we would insist (relatively uncontroversially) that a person is committed to an answer to (a) certain question(s) just in case that person has a certain attitude, but we would allow that one might have the attitude without having engaged in the activity of settling the question. Thus, if one does settle the question(s) for oneself, one exercises control over the attitude. However, someone might have that attitude, and so be committed to an answer to the question—someone might have the "normative status"—without ever having settled the question, and perhaps without ever having exercised any agency with respect to the attitude. On this interpretation, the attitude embodies one's answer to a question, but it does not, thereby, embody an exercise of agency.

On a second, more radical, interpretation, we would not allow (in persons) being committed to an answer to a question to part company with having settled that question. Rather, we would insist that, if you are committed to an answer—that is, if you are open to those questions and criticisms that would be satisfied by reasons that (you take to) bear on a question or set of questions—then you must have settled that question. On this interpretation, the "normative status" cannot appear apart from an exercise of agency. Rather, an exercise of agency (viz., the agency at work in settling a question for oneself) incurs the commitment, in each case. (This more radical interpretation would simplify the complex conjunction given above as a definition of "embody an answer to a question.")

Recalcitrance

One might first object that we sometimes settle certain questions without thereby altering either our attitudes or the commitments they entail and are entailed by. Thus, one might think, evaluative control is at best not entirely reliable, and, moreover, I should revise or qualify my claim that, if you settle for yourself this or that question, you therein form this or that attitude. But this claim was, it seems, the main motivation for claiming that we exercise evaluative control.

My reply to this objection will seem, at first, cheap: Insofar as you have in fact settled a question, to that extent you do change your commitments (i.e. the questions and criticisms to which you are answerable), and insofar as you have not changed your commitments, you have not in fact settled that question. But, insofar as you have changed your commitments, you have formed or altered the associated attitude. So, if you have in fact settled a question, then you must have formed or altered the associated attitude.

This reply may seem cheap, because may seem that I am simply defining "settling a question" so as to ensure my claims are correct, against an obvious, intuitive problem. So I will try to show that, even in the problematic cases, my seeming stipulation is plausible.

This more radical interpretation will obviously require positing an exercise of agency in a surprisingly wide range of cases, and so, one might think, either will be wildly implausible or else will require an objectionably deflationary account of agency—agency will be attributed wherever we find an attitude with a certain sort of "normative status," regardless of whether we find, there or in the agent's history, any discernable mental processes or activities that we could independently identify as an exercise of agency which we might associate with that attitude. Though I am currently inclined to think that we should prefer the more radical interpretation and accept the unusual understanding of agency it entails, defending this (initially implausible or deflationary) choice will require considerable work, and I will not here undertake the task. Rather, I will note that, on either interpretation, one exercises agency over certain attitudes in settling questions for oneself. In the text I concern myself with this weaker claim, which raises enough worries for present discussion. (For an excellent discussion of some of the difficulties that might plague the stronger claim, see Matthew Boyle, "Making up Your Mind," (in progress).)

By adopting my seeming stipulation, we preclude the possibility of settling a question without therein changing one's commitments and so one's attitudes—but such cases seem not merely possible but actual. We can identify them because we are sometimes able to identify the settling of a question apart from the presence of certain attitudes: you can, e.g., settle a question by engaging in a conscious, overt process of deliberation on that question, and coming to a conclusion.[5] But, of course, you might deliberate and come to a conclusion that is at odds with the attitudes you continue to hold. You might believe not *p*, or intend not to ϕ, and then deliberate about whether *p*, or whether to ϕ, and reach a positive conclusion. You might nonetheless continue to believe not *p*, or intend not to ϕ. And this, one might think, shows that you can settle a question without changing or your commitments or controlling your attitudes.

I agree with all but the last claim. If you have, in fact, concluded that *p* (e.g), then it seems to me that you will, at least for a moment, incur the commitments associated with believing *p* and, therefore, that you do, at least for a moment, believe *p*—perhaps despite the fact that you also continue to believe not *p*. Thus, in the problematic situation, either you have, upon reaching your conclusion, arrived at the conflicted and difficult state of believing *p* and also believing not *p,* or else you are momentarily waffling in your beliefs about *p*. Saying either seems to me more plausible than saying that you have somehow come to a conclusion without changing your commitments and therefore your attitudes.[6]

[5] It is important that this is not the *only* way that you can settle a question. But it is one way.

[6] Better, it seems, to locate the difficulty in the particular thinking subject, who is conflicted or inconsistent, than to allow that one can settle a question without incurring the associated commitments or to allow that a person's commitments and attitudes can part ways. But someone might disagree about this last claim. I consider such disagreement in the next footnote.

Of course, if we accept either of the preferred descriptions, it will be true that, in coming to the conclusion that *p*, you *have* exercised a kind of control over your mind—you formed a belief that *p* in settling for yourself (positively) the question of whether *p*. What you have not done is exercise control over your belief that not *p*. In order to control *that* belief, it seems you will have to find a way to keep yourself consistent—but, importantly, keeping yourself consistent is not required for an exercise of evaluative control.

This response may seem disappointing. To really control your attitudes, one might think, you should be able to target a specific belief—the belief that not *p*, say—and see to it that *that* belief changes, when you settle a question, if you think it should. Short of this, one might think, what I am calling evaluative control does not deserve to be called a kind of control. I will take up this kind of worry next.[7]

The Paradigmatic Features of Agency or Control

So, I hope it relatively plausible that, as you settle for yourself certain questions, you therein, *ipso facto*, form or revise certain attitudes. As you make up your mind about

[7] The reply might disappoint in another way: the example may seem to call into question my claim that a person has the attitude just in case that person is rightly open to certain questions and criticisms (i.e., committed). Perhaps, one might think, a person can be (momentarily) subject to criticisms (as a result of settling a question) without therein (momentarily) believing, or have a belief without incurring the typical commitments. I resist this position in part because it seems to me that an attitude, in a person, that does not support the relevant commitments will not be a belief or intention, but rather a thought, fixation, wish, or inclination—something less than a person's belief or intention—and in part because it seems to me that one cannot rightly be subject to the relevant questions and criticisms unless one in fact believes or intends.

However, perhaps surprisingly, I suspect that granting this objection, and so allowing commitments and attitudes to part ways, would complicate but not entirely upend the view here presented. On the more complicated version of the view, evaluative control would be exercised over the commitments, which would in turn bear some relatively close but not necessary connection to the associated attitudes. Thus, on such a view, one will not only have to keep oneself consistent, somehow, by means other than evaluative control, but also keep one's attitudes in line with one's commitments, somehow. (While this seems a possible view, I would prefer to keep the commitments more clearly associated with some psychology; some such association seems inevitable, and belief seems a good candidate for the job.) I devote myself, in the main text, to what I think is the more pressing and illuminating objection.

what is true, or what to do, you therein, in some sense literally, make up your mind—you create or constitute, form or revise, your beliefs and intentions. These attitudes, one might say, just are your take on their object, and so, when you change your take on their object, you therein change these attitudes.

While such simple reflections lead naturally to the thought that a thinking subject controls its thoughts (or, at least certain of its thoughts) as it thinks them, there is some reason to resist calling this a form of control, because some of the most salient features of the paradigmatic instances of agency or control are lacking, in this case.

Agency is paradigmatically exercised in ordinary intentional action. *Control* is typically exercised by some subject on some object, where, paradigmatically, the subject has some intentions about the object and controls the object by successfully executing those intentions. Thus it seems that one paradigmatically exercises agency or control by (successfully) executing one's intentions with respect to an action or object. Thus we are led to expect certain features of any exercise of agency or control: we expect exercises of agency or control to display both a certain kind of *voluntariness* (in one sense of that difficult word) and, relatedly, a certain kind of *reflective distance* or *awareness*. But evaluative control displays neither of these features. The forming and revising of beliefs and intentions is not voluntary nor does it require the same kind of reflective distance or awareness.

To illustrate, consider first ordinary intentional actions (such as getting some lunch or managing one's finances). When we intend to do something, it seems we have, in some sense, settled for ourselves positively the question of whether to do that thing—a question that represents the action, under some description. In settling that question, we form an

intention, which intention we will, if all goes well, execute in intentional action. Moreover, we can settle the question of whether to ϕ, like any question, for any reason(s) we take to bear convincingly on it—or perhaps for no reason at all. I can, e.g., decide to get some lunch for any reason(s) that I take to settle the question of whether to do so. Thus I will say that ordinary actions are *voluntary*, in the following, somewhat technical, sense: we can, for any reason that we take to count sufficiently in favor of the action (or perhaps for no reason at all), settle the question of whether so to act, therein intend so to act, and, providing as all goes well, execute that intention in action.

A certain kind of reflective distance or awareness goes hand-in-hand with this kind of voluntariness: if we form our intentions by settling for ourselves a question that represents our action under some description, then it seems that our action is, in some sense, an object of our thought—in a way that, e.g., the unforeseen consequences of our actions are not.[8]

The same features appear in the paradigm cases of control over ordinary objects—over cups and cars and computers. Since we control these objects by forming and successfully executing intentions with respect to them, it seems that the ordinary objects of ordinary control are, in the paradigmatic cases, represented or implicated in the

[8] It may seem problematic to move from the claim that one is committed to an answer to a question that represents the action to the claim that one represented that action in thought. It seems I have moved from a claim about the criticisms to which one is rightly vulnerable to a claim about what sort of events have occurred in one's mind. While I am tempted to make such moves, I do not think this one is strictly necessary for the point at hand. I think it clear enough that, if we act intentionally, the action we intend is (paradigmatically?) represented to us in a way that unforeseen consequences of our actions are not represented. This will contrast with the attitude themselves. It seems that our beliefs, e.g. (that is, our own states of mind) are not represented, as we form them.

question we settle for ourselves, and, again, we can settle that question for any reason(s) we take to bear convincingly on it.[9]

Thus, in the paradigm cases of agency or control, that over which we exercise control or agency—whether an action or an ordinary object—is, in some sense, a part of the content of our thought, in a way that, e.g., the unforeseen consequences of our actions are not. In the paradigm cases, there is a certain familiar reflective distance between the subject who controls and the object that is controlled, or between the agent and what the agent affects (or effects). We exercise agency or control, one might say, when we are the cause of our own representations—the cause of that which we represent as to be done.[10] Moreover, in this "reflective distance" we encounter a kind of voluntariness: in reflecting upon the action or object of control, we can decide to do that which we have in mind to do for any reason we take to settle the question of whether to do it.[11]

Evaluative control display neither of these familiar features: the objects of evaluative control (beliefs and intentions) need not stand at a reflective distance in our thought as we

---

[9] This claim that an ordinary object of control is represented in the question settled will be more controversial than the claim that the intentional action is so represented. After all, it seems you will control your pen in executing an intention to write a note. It does not seem that the pen is represented in the question of whether to write the note. Still, I think it plausible to say that, at some point, your intention to write the note will involve some representation of the pen, since your use of the pen was not unforeseen. Perhaps your intention to write a note leads to an intention whose content has something to do with your pen. There are various ways to understand such a "nesting" of intentions. For discussion of related issues see Michael E. Bratman, *Intention, Plans, and Practical Reason* (Cambridge: Cambridge University Press, 1987) and G. E. M. Anscombe, *Intention* (Oxford: Blackwell Publishing Co., 1957), 37–47.

[10] Notably, Kant defines the capacity for desire as "the capacity to be by means of one's representations the cause of the objects of those representations" (*Metaphysics of Morals* 211 and *Critique of Practical Reason* 5:10). Many seem to find being the cause of one's representations a necessary, but not sufficient, feature of agency: we are agents, they think, when we not only cause what we have, in some way, represented, but when we do so intentionally—when cause something we have represented because we have in some way decided to cause it. On such a picture, to be an agent is to be able cause the objects of your representations voluntarily: to be able to exercise a kind of executive capacity over which of your desires is actualized (over which of your representations are the cause of that which they represent).

[11] It may be worth noting that the discretion here does not include the ability to do something even when you are convinced that you have sufficient reason not to do it.

exercise this form of control over them—we need not represent our beliefs and intentions in the way we represent our actions. Nor is their formation or revision voluntary in the way ordinary intentional actions are voluntary—we cannot form or revise or maintain or create them for any reason we take to count sufficiently in favor of so doing, but only for reasons proper to them: only for reasons that we take to show the belief true or to bear on whether to perform the action intended.[12] Nonetheless, I think we should grant that we do exercise a kind of control or agency over our attitudes as we settle for ourselves the questions they embody.

I will elaborate on these claims in a moment, while defending the thought that evaluative control should be counted as a kind of agency. But first, to avoid confusion, it will help to note that, on the account presented here, many exercises of mental agency will be instances of what might be called *mental actions*. That is, many exercises of mental agency will share the structure and display the familiar features of ordinary intentional action. So, e.g., you might call to mind where you put your keys, try to remember the last time you visited your sister, rotate an object in your imagination, or picture your living room walls a different color. So long as such imaginings and rememberings are intentional, they can be classed, on the account here presented, with ordinary actions like raising your right hand or getting some lunch.[13] No doubt there are many interesting and important questions about the various forms of mental action, but I

[12] Another important dissimilarity: if I am right about the relation between settling a question, incurring a commitment, and having an attitude, then the relation between settling a question and forming the attitude that embodies one's answer is not causal, but rather something like conceptual or constitutive. There is no possibility of things going wrong, between one and the other.

[13] Of course, it may be that you remember something, or that something appears in your imagination, unintentionally—the thought comes unbidden, so to speak. I presume that these mental goings-on need not be treated as instances of mental agency, and so leave them aside.

will not address them here. Rather, I will simply class mental actions with other ordinary actions, and contrast them with the agency at work in the formation and revision of such attitudes as belief and intention.

Doing without the Paradigmatic Features

Why should we allow that what I am calling evaluative control deserves to be thought of as a kind of agency? There is much to be said, but I will confine myself to some brief remarks.

First, to avoid verbal dispute, it should be granted that one might well reserve the word 'agency' for those activities that do display the familiar features of voluntariness and reflective distance. Such usage would be unobjectionable, so long as it does not invite the thought that anything lacking the distinctive features must be a kind of passivity, or something merely acted upon. Thus, I would insist that some title should be granted to evaluative control (perhaps we could call it a kind of "activity") that prevents its exercise from being grouped with those things that merely happen to one and prevents its outputs—the attitudes I claim one forms or revises by means of its exercise—from being grouped with those things that one can affect only by acting upon them.

It seems to me plain that we need some additional category of agency or activity—one that does not share the characteristic features—in order to accommodate the agency we exercise over our own intentions. We have already granted that the formation and revision of intention is not voluntary. In fact, I have argued elsewhere that it could not be.[14] Nor, it seems to me, need intentions be represented in thought as one forms or

[14] See, e.g., Pamela Hieronymi, "Controlling Attitudes," *Pacific Philosophical Quarterly* 87, no. 1 (2006), Pamela Hieronymi, "Responsibility for Believing," *Synthese* 161, no. 3 (2008).

revises them. But if forming and revising an intention does not display the familiar features of the paradigmatic exercises of agency, then it seems that these features could not be essential to agency—since it seems we must exercise some form of agency in forming and revising our intentions, if we exercise agency at all.

One might wonder why I should claim that intention is not, and could not be, voluntary in the sense at issue. I will first present a pair of cases that I hope will lend the claim some intuitive support and then briefly sketch the argument I have given elsewhere.

Consider, first, a case that seems to suggest that intention *is* voluntary: Suppose an experimental psychologist with an "intention-detector" offers you a small sum for intending to drink some water. It seems you can decide to form the intention and earn the money. Thus it seems that intending is like raising your right hand—something you can do on command, as a so-called "basic action." More to the point, it may seem that intending is voluntary in the way ordinary action is voluntary: it may seem that you can decide to intend for any reason that you think shows intending worth doing.

Now suppose instead that the psychologist would like to see register, on her machine, an intention to jump from the third-story window, and she offers you the same small sum for forming that intention. You might well think the small sum is well worth *intending* to jump (no harm, you think, in simply intending). But, of course, you will not intend to jump, and so will not earn her reward, unless you are committed to *jumping*—unless you have settled for yourself positively the question of whether to jump. And the small sum is not, you think, reason enough to settle that question. But if you do not think the sum reason enough to settle the question of whether to jump, then (assuming that you have no other reasons for jumping and some reasons not to jump) it seems that you cannot

respond to her offer by intending to jump, and so cannot earn her money. So it seems you cannot, in this case, intend for reasons that you think count sufficiently in favor of intending.[15]

What prevents you from earning the reward, in the second case? I suggest it is the fact that one intends only if one is committed to an action but one cannot become committed to an action by finding convincing reasons that one only takes to show the intention good to have. That is to say, one cannot become vulnerable to questions and criticisms that would be satisfied by reasons that (one takes to) bear on whether to ϕ by finding convincing the reasons that one does *not* take to settle this question, but which one rather takes to settle the distinct question of whether the intention to ϕ is good to have. Thus, one cannot form an intention for any reason that one takes to count sufficiently in favor of intending; one can only form an intention for reasons one takes to settle the question of whether to act. In contrast, one can act for any reason one takes to

[15] The case is science-fictional, but it need not be. There are plenty of everyday cases in which a reason for an intention is not reason enough to act. Perhaps it displeases you that I do not intend to finish my work by tomorrow. And perhaps you would be satisfied simply knowing I intend, regardless of whether I actually finish. And perhaps I am generally happy to house mental states that please you. Still, I will not be able to intend to finish, in order to please you, unless I also take pleasing you to be reason enough, not just to house the intention, but to finish.

These cases are, of course, variations on Kavka's Toxin Puzzle, found in Gregory Kavka, "The Toxin Puzzle," *Analysis* 43 (1983). The case of intending to jump out the window is unlike Kavka's puzzle, in that, in Kavka's puzzle, the reward for the intention is well worth performing the action (in this way Kavka's case is like the case of drinking the water). The case of drinking the water is unlike Kakva's puzzle in that the action is performed immediately and carries no disincentive. I consider Kavka's puzzle in a lengthy footnote in "Controlling Attitudes."

Niko Kolodny points out that, in any such example (science-fictional or no), any reason against acting will also be a reason against intending so to act, since your intentions are likely to lead to action. Thus, he thinks I have not yet provided a case in which you have sufficient reason to intend though you lack sufficient reason to act, and so he remains unconvinced of my claim that you cannot intend for any reason that you take to count sufficiently in favor of intending. For all I have said, it may still be the case that you can intend for any reason you take to count sufficiently in favor of so doing. I grant that the examples do not establish the claim. For further treatment of Kolodny's objection, see footnote 16.

count sufficiently in favor of acting. Thus, intending is not voluntary in the way ordinary action is.[16]

Why, then, can you earn the reward in the first case? We can give the following interpretation: When the psychologist offers you the small sum to intend to drink the water, you can take the offer to be reason enough to settle the question of whether to drink, therein decide to drink, and so intend and earn the money. But you cannot do the same, in the second case, because you do not think the small sum is worth the jump.

If this treatment of these cases is correct, then intending is not voluntary in the way an action is: you cannot form, revise, or maintain an intention for any reason you think counts sufficiently in favor of forming, revising, or maintaining it. Rather, you can only form, revise, or maintain an intention for reasons that you take to settle the question of whether to act. But you can act for any reason you take to count sufficiently in favor of acting. And so it seems that intention is not voluntary in the way that ordinary action is.[17]

---

[16] For reasons that one takes instead to show an intention good to have (which one does not take also to show the action worth doing), one *could* form an intention *to bring it about* that one forms the desired intention—one could, by finding convincing reasons that one takes to show an intention good to have, commit to the action of bringing that intention about. But, again, one need not make such managerial commitments in the case of ordinary action: ordinarily one need not form an intention to bring it about that one acts; one simply forms an intention to act, and executes that intention in the action. Thus, again, intention is not voluntary in the way that ordinary action is. The argument of this paragraph appears in both "Controlling Attitudes" and "Responsibility for Believing."

[17] Sometimes, in response to this sort of argument, people insist that you can form, revise, or maintain an intention for any reason you take count sufficiently in favor of doing so, but add that the question of whether to form, revise, or maintain an intention to ϕ is "transparent to" the question of whether to ϕ—that these questions must be answered by the same set of reasons. In this case, asking yourself whether to intend to ϕ seems simply to be a (somewhat sophisticated, reflective) way of asking yourself whether to ϕ. It is sophisticated or reflective (at least) in that it brings to one's mind the fact that, if one decides to ϕ, one will, therein, intend to ϕ. (This is closely related to Kolodny's objection, above.)

(Richard Moran developed an account of transparency in his investigation of self-knowledge. Notably, Moran thinks the question of whether I believe *p* (e.g.) is, insofar as I am rational, transparent to the question of whether *p*—i.e., these questions will be settled by the same reasons. See Moran, *Authority and Estrangement: An Essay on Self-Knowledge*. Nishi Shah considers a transparency thesis closer to the one here considered in Nishi Shah, "How Truth Governs Belief," *The Philosophical Review* 112 (2003).)

Once we grant, however, that forming an intention is not voluntary, it seems that we cannot require that every exercise of agency be voluntary: because it seems that the forming of an intention must be an exercise of agency, if anything is.

One might grant that voluntariness of the sort specified is not essential to exercises of agency, but hold out for the other familiar feature: the reflective distance or awareness. Though many seem to find this feature important, it seems to me inessential. I will too briefly suggest why.

Notice, though, that if we insist that the question of whether to intend to ϕ can be settled only by reasons that bear on whether to ϕ, it seems that we have given up the thought that intending is voluntary in the way that ordinary action is voluntary. An ordinary action is voluntary in that it can be done for *any* reason that one takes to count sufficiently in favor of so acting. But, on the interpretation just given, intending to ϕ cannot be done for any reason one takes to count sufficiently in favor of so intending. Rather, one can decide to intend to ϕ only in those cases in which one can decide to ϕ.

One might, at this point, return with Kolodny's objection. Recall that Kolodny doubted that there would be cases in which one has sufficient reason to intend to ϕ but lacks sufficient reason to ϕ, because ϕ-ing is a(n obvious) consequence of intending to ϕ. So, the bad effects of jumping show that you do not have sufficient reason to intend to jump. Following this line of reasoning, one might think that intending might be voluntary after all: maybe you *can* intend to ϕ for any reason that counts sufficiently in favor of so doing. It just turns out that you will have such reasons only in cases in which you also have sufficient reason to ϕ.

But even if one established that the only considerations that in fact count sufficiently in favor of intending are those that count sufficiently in favor of acting, and so established that a person can intend to ϕ for any reason that (in fact) counts sufficiently in favor of so doing, one would not thereby undermine my claim. My claim is that, while you can (intend to act, and, providing all goes well) act for any reason that you take to count sufficiently in favor of so acting, you cannot intend to ϕ for any reason that *you take to* count sufficiently in favor of doing intending. So, to undermine my claim, one would have to establish, not just that the only reasons for intending are those that are (in fact) reasons for acting, but that no one could *take* reasons to count sufficiently in favor of intending without also *taking* them to count sufficiently in favor of jumping (Shah is aiming at something like this position, with respect to belief, in his Shah, "How Truth Governs Belief."). But it seems possible that someone might take that view, even if it is mistaken. So, suppose someone (perhaps mistakenly) thought that the small sum counts sufficiently in favor of intending to jump, without taking it to count sufficiently in favor of jumping. My claim is that such a person cannot intend for the reasons that she takes to count sufficiently in favor of intending, though she could (providing all goes well) jump for *any* reason that she takes to count sufficiently in favor of jumping.

To put the point another way: you will intend to ϕ only if you are committed to ϕ-ing, and (if you commit to ϕ-ing for reasons) you can only commit to ϕ-ing for reasons that you take to settle the question of whether to ϕ. But you might (perhaps mistakenly) take certain considerations to show intending to ϕ worth doing, which you do not take to show ϕ-ing worth doing. You will not be able to intend for these reasons (though, as noted, you may be able to bring it about that you intend for those reasons). In contrast, you can (intend to ϕ and, providing all goes well) ϕ for any reason you take to show ϕ-ing worth doing.

Consider, again, intention. Insisting that an exercise of agency must involve the characteristic reflective distance or awareness requires us to say that the forming of an intention was not an exercise of agency unless the agent had some thought about or awareness of that intention—indeed, unless the agent had a thought about or awareness of the intention *of the sort* characteristic of our thought about or awareness of our own actions or the ordinary objects we control thereby.[18] But it seems to me implausible to claim that we are typically, or even very frequently, thus aware of or reflective about our own minds (as opposed to the actions we intend or the ordinary objects we control). Further, it seems that a lack of such awareness of our minds does not distract from the agency we exercise in acting.

Suppose, to be fanciful, that someone is part of a psychological study, in which she is taking a drug that will make her nauseous if she forms an intention to stay up late. She is now, under stress, trying to figure out how to finish all the projects she must accomplish by the end of the week. In trying to work out this practical problem, she plans to stay up late tonight, but she does so while forgetting not only that forming such an intention will make her nauseous, but also unmindful, even, of the fact that she has just formed an intention—unmindful of the fact that she has just changed her psychology. It seems to me that, in this case, not only the bad effect but even her intention itself is an unforeseen consequence of her attempt to solve her practical problem. And yet, for this lack of

[18] Some will want to insist that an intention occurs in its own content, and so think that they have secured for intention the paradigmatic feature of ordinary action. While there may be other reasons for insisting that an intention occurs in its own content, I doubt that this strategy can plausibly gain for intention the sort of awareness that is characteristic of ordinary action or control over ordinary objects.

A full treatment of this claim will obviously require some account of how we are aware of our actions and the objects we thereby control. I have given some indication of my account of this, above: we settle a question that represents our action, under some description.

awareness of her own mind, her decision to stay up late seems no less an exercise her agency.[19] And thus it seems to me that we can exercise agency with respect to our intentions even when we are not aware of them in the characteristic way in which we are typically aware of our intentional actions.[20]

Allowing that we can be agents with respect to our attitudes even when we are not aware of or reflective about them goes against a powerful intuition. It seems very odd to think that we can be exercising our agency—and do so normally and well—by creating something that we did not intend to create and that remains, so to speak, out of our own view, behind our back, or off-stage—something that may well be "unforeseen." But once we notice that, whenever we make a decision or come to a conclusion on some topic, we therein make something true of our own minds (namely, that we have decided or concluded); that we can do so without having any intentions about our own minds; and

[19] I am very grateful to Yannnig Luthra for his thoughts on this example, and for pressing for clarification.

[20] So-called "Freudian slips" provide a different kind of example in which one intends without awareness of one's intention; these are sometimes taken to show that "full-blooded" agency requires some awareness of one's own intentions and/or motivations (see, e.g., J. David Velleman, "What Happens When Someone Acts?," in *Perspectives on Moral Responsibility*, ed. John Martin Fischer and Mark Ravizza (Ithaca: Cornell University Press, 1993).) But it should be noted, I think, that in such cases one is (also) unaware that one is doing the action in question (either at all or, at least, under the description under which the action is a "slip"). So, even if it is granted that these are not cases of full-blooded agency, this might show only that full-blooded agency, when exercised in action, requires an awareness of the description under which one is in fact acting, not that it requires an awareness of one's state of mind, or of one's motivations, or an awareness that, as one decides to act, one is making certain things true of one's psychology.

We should wonder why awareness of one's intention would be thought to make one's agency over one's action more full. There is, of course, one way in which such awareness enhances one's agency over one's action: if one is aware of the fact that, in deciding to act, one will therein change one's state of mind, then one is more fully aware of both the possible reasons for and possible consequences of one's action. Being so aware, one can, e.g., decide to drink in order to form an intention and earn the small sum, or decide against ϕ-ing in order to avoid the bad effects of an intention to ϕ. But this is just to say that an awareness of one's own mind can enhance one's agency in acting in just the way that any further relevant information can: I am, in this sense, more fully an agent anytime I am more fully aware of all my options, or all my possibilities—and more fully an agent the less that remains unforeseen. I would readily grant that one's agency in the case at hand is less than full, in this sense. But this can be granted without damage to the point: one exercises agency of an ordinary, non-defective sort over one's action, even when one does not have in mind one's mind.

that, even if we were to turn our attention to our own minds and to make decisions and come to conclusions or form intentions about it, we would, in so doing, create a higher-order set of attitudes with the same disquieting features, we might start to think that this intuition is simply a bias born of our familiarity with our agency as exercised in our actions and over ordinary objects. I believe we should go without it.

I hope, then, that I have at least suggested why we might allow that evaluative control is a form of control or agency, despite the fact that it lacks the familiar features. I will now, as promised, consider how managerial control can seem to involve an exercise of evaluative control, at two distinct points, and how certain familiar, complex exercises of agency over our own minds can be more clearly understood as so composed.

Managerial Control and Its Decomposition

Consider, first, the great variety of methods by which one might manage or manipulate one's own beliefs or intentions—the variety of methods by which one might take action so as to affect one's beliefs and intentions according to one's purposes:

Most bluntly, you might bring it about that you believe p or intend to ϕ by doing something that affects your brain in a way that is likely to have this effect. If, e.g., you want to believe that your friend has never betrayed you, you might induce in yourself amnesia about the relevant stretch of shared history. If you want to believe that this or that is not so worrisome, you might take some anti-anxiety medication. Perhaps, at some point in the future, we will be able to induce particular beliefs or intentions directly, by taking a pill or stimulating the brain. Perhaps hypnosis produces a similar effect.

At the opposite extreme, you might bring it about that you believe p simply by changing the world so as to make p obviously true. As pointed out by Richard Feldman,

if you want to believe the lights are on in your office, you can get up and throw the switch.[21]

Less radically, you can also manage your own attitudes by taking steps that you can predict will provide you with convincing reasons for the answer embodied in the attitude. So, again, if you want to believe that your children arrived home safely through the storm, you might call them and thereby provide yourself with convincing evidence that they have. If you want to be sure that, tomorrow, you will do the right thing, you might tell your friends about your plans, today.

(There will be some difficulty, of course, if you believe that you have provided yourself with skewed or unfair evidence for *p*—because this belief will make the evidence less compelling in your own eyes, and so make it less likely that you will conclude that *p* on the basis of it, and so make it less likely that you will successfully bring it about that you believe *p*. As noted earlier, while you can bring it about that someone else believes by providing that person with reasons that you do not, yourself, find convincing, you cannot do the same to yourself. Self-deception is notoriously harder than deceiving others.)

Somewhat more subtly, you might manage your own attitudes, not by providing yourself with new reasons, but by convincing or persuading yourself that the reasons at hand support an alternative conclusion. You might intentionally direct your attention in certain ways, or provide yourself with alternative interpretations of your situation, or persuade yourself to "see things differently," or take steps to convince yourself that your own previous response is unjustified or that an alternative response is equally justified.

[21] Richard Feldman, "The Ethics of Belief," *Philosophy and Phenomenological Research* 60 (2000): 671–2.

Or you might take steps to keep your attention focused on the reasons you already find convincing, which you predict you will be tempted to overlook in the future: if you want to strengthen your dieting resolve, you might post a picture on the refrigerator door.

In any of these ways, then, you might take steps to bring it about that you believe *p* or intend to ϕ, in much the same way that you might take steps to bring it about that someone else believes (or, for that matter, in the same sort of way that you might bring it about that your living room walls are pale green): you take action designed to bring about that end, subject to the ordinary sorts of limitations one always encounters in trying to effect changes in the world.

These cases in hand, note that a successful exercise of manipulative or managerial control over one's attitudes will involve, at two distinct points, a commitment to an answer to a question, and so, it seems, might involve two distinct exercises of evaluative control.

First, if we assume that an exercise of managerial or manipulative control is intentional (as it seems it must be, to earn the title[22]), then, when one exercises managerial or manipulative control over one's attitudes, one will intend so to manage or manipulate one's attitudes, and so will be committed to a positive answer to the question of whether to do so. This commitment would seem to be the result of having settled for oneself the question of whether to manage or manipulate—the result, that is, of an exercise of evaluative control.

[22] This should not confuse: an exercise of managerial or manipulative control must be intentional, to qualify as control, despite the fact that an exercise of evaluative control need not be. Managerial or manipulative control is a matter of acting so as to affect something according to one's purposes. If one acts so as to affect something according to one's purposes without intending to, it seems wrong to say that one has exercised control over that thing.

Second, if you succeed in your exercise of managerial or manipulative control, you will bring about an attitude that embodies your answer to a question. Thus, if you succeed, you will have brought it about that are committed to whatever answer is embodied in the attitude.[23] While there is room for disagreement about whether, in bringing about this commitment, you will have brought about an exercise of agency, it should be granted that, at least in certain cases, one can bring it about that someone (perhaps oneself) believes or intends by bringing it about that that person exercises his or her agency in a certain way.

So it seems that any successful exercise of managerial or manipulative control will require a commitment to an answer to a question at two distinct points: one will be committed to a positive answer to the question of whether so to manage or manipulate one's attitude and one will be committed to whatever answer is embodied in the attitude successfully managed or manipulated.[24] Either of these might involve an exercise of evaluative control.

[23] Of course, one might bring about this second commitment—the commitment embodied in the target attitude—either honestly, so to speak, or dishonestly. This accounts for the continued use of the cumbersome disjunction, "managerial or manipulative control." As we saw, you might bring yourself to believe *p* by making *p* obviously true, conducting a fair investigation, or providing for yourself evidence, or you might take steps that produce incentives that ensure that you will intend to ϕ, or persuade yourself to take up another, equally reasonable, point of view on *x*. If you bring about the commitment by any of these "honest" means, it will seem right to say you *managed* your attitude, or that you exercised *managerial control* over it. But we are not restricted to honest effort. You can bring it about that your attitudes change in ways that produce irrationality or require some kind of self-deception or amnesia. In such cases it will seem right to say that you have *manipulated* your attitude, or that you have exercised *manipulative control* over it. I suspect the distinction between management and manipulation will be hard to draw sharply; happily, we need not draw it sharply, for present purposes. We can simply note that all of these methods belong to a genus: they are ways of acting so as to bring it about that you form or revise or maintain some attitude, which attitude itself embodies your answer to a question.

[24] Or, of course, if one successfully rids oneself of an attitude, one will then cease to be committed to the answer it embodies.

Thus it seems that these two forms of agency can display a characteristic division of labor in an exercise of managerial or manipulative control: Perhaps you decide to manage something—in the cases at hand, some attitude of yours, which we will call the target attitude. That decision itself constitutes an exercise of evaluative control with respect to an intention—in deciding to manage the target attitude, you form an intention about it. That intention is executed in a managerial or manipulative action, aimed at changing the target attitude. (Of course, the action here may be mental: it may consist of directing your attention in certain ways, calling to mind certain facts, presenting yourself with an alternative interpretation, or vividly imagining certain outcomes.) Insofar as your managerial or manipulative actions succeed in their aim, you will bring it about that you are committed to the answer(s) embodied in the target attitude. This might involve bringing it about that you settle the relevant question(s) in the relevant ways, and so might involve inducing or influencing the conclusion you come to on some question—that is, it might involve inducing or influencing an exercise of evaluative control.

If we allow that these two forms of control can thus work in tandem, it seems that an exercise of evaluative control can be induced or produced by an exercise of managerial control and that an exercise of evaluative control can initiate each exercise of managerial or manipulative control. Some will find both these claims unsettling or disorienting. The first will seem unsettling because it can seem that exercises of agency should not be the sort of thing that can be induced or brought about or manipulated. But to so insist is to deny not only some of the most important forms of self-management but also some of the most obvious forms of moral wrongdoing (those that involve the manipulation of

another's will).[25] The second claim will seem unsettling because it will seem that whatever initiates an exercise of control should display the familiar features of agency: it should be voluntary and should involve reflective distance or awareness. Though I have already suggested why I think we need to allow a form of agency that does not display these features, I suspect that my reflections may not unseat the strong intuition. I will close, then, by briefly considering one popular alternative model of our agency over our minds, one that preserves the familiar features.

REFLECTIVE CONTROL

Many philosophers are drawn to a class of accounts of our agency over our minds that I will group together under the head *reflective control*. On such accounts, we exercise agency over attitudes like belief and intention by reflecting critically upon them and determining for ourselves whether they are justified.[26]

Reflective control is attractive, at least in large part, because it seems to preserve the paradigmatic features of ordinary agency. After explaining how it does so, I will suggest that reflective control is difficult to model—it is difficult to understand just how it works. I will briefly mention some ways in which it has been modeled and suggest why I find

[25] In his presidential address, Rogers Albritton might seem to suggest that your will cannot be manipulated. But I think he is in fact making a different point: that you cannot, as a conceptual matter, be made to will something against your will—if you are made to will it, you then will have willed it. While this is doubtlessly true, it is hardly a defense against manipulation. See Rogers Albritton, "Freedom of the Will and Freedom of Action," in *Free Will*, ed. Gary Watson (Oxford: Oxford University Press, 2003).

[26] A nice recent discussion of various proponents of reflective control can be found in David Owens, *Reason without Freedom: The Problem of Epistemic Normativity* (London: Routledge, 2000). A slightly different set of accounts, also deserving of the name, have it that we exercise agency over our attitudes by determining for ourselves whether we want to have them, or whether they make sense to us. See, e.g., the papers collected in Harry Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988) and J. David Velleman, *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000).

them dissatisfying before suggesting that we might be able to construct (what seems to me) a more satisfying account by employing the accounts of evaluative and managerial control I offered above (together with an assumption that a proponent of reflective control must also employ). Of course, the account I have offered abandons the thought that our agency over our minds must display the paradigmatic features of agency. Thus, if we use it to model reflective control, it seems we might give up the thought that reflective control is the primary way in which a rational agent exercises her agency over her own mind.

To begin, we need a clearer understanding of the phenomena I am calling reflective control. Many have been powerfully struck by the fact that we can change our own attitudes simply by reflecting on whether they are justified. It is, indeed, a striking fact. After all, reflecting on the justification of this or that does not typically alter the object of the reflection. I may, e.g., reflect on some belief or intention of yours, and come to the conclusion that your belief is unjustified or your intention unsound, without thereby having the least effect upon your belief or intention. Moreover, I may even communicate my reflections to you, without thereby changing your attitudes—and neither of us need be, for the ineffective exchange, in any way irrational. We may simply, reasonably, disagree. But if you find, upon reflection, that one of your own beliefs is unjustified or one of your intentions is unsound, then, often enough, that reflection itself seems sufficient to undermine the attitude. Of course, it does not always do so—sometimes you can find yourself in the inconsistent position of believing something you also, in reflection, have determined unjustified, or intending to do something you also think a bad idea. But in such a case you are, it is said, in some way irrational. Thus it seems, insofar

as you are rational, reflecting upon whether your attitudes are justified will, itself, change them.

Note how reflective control seems to share the paradigmatic features of ordinary agency: we can intentionally, for any reason we see fit, decide to reflect upon whether our attitudes are well grounded. Reflecting upon one's own attitudes can be voluntary, and can be done for a purpose. Often enough it is done for the purpose of ensuring that one's attitudes are justified. (Insofar as it is done for that purpose, then, insofar as one is rational, it will achieve its end.[27]) Further, when we reflect on our attitudes, we certainly stand at some reflective distance from them. When we determine whether some attitude of our own is justified, we come to a conclusion about *it*—about the object of our reflection. Still, even though reflective control displays the paradigmatic features of agency, changing your attitudes by reflecting upon their justification seems quite unlike acting upon them in a merely managerial or manipulative way—quite unlike acting upon them in the way we act upon ordinary objects. These facts make attractive the thought that it is reflective control, rather than evaluative control, that provides the best model of our distinctive agency over our own attitudes: we are agents over these attitudes (or perhaps most fully agents over them), not when we simply reflect on and come to conclusions about their content, but when we reflect on and come to conclusions about whether they are justified.

Before adopting this model, I think we need to better understand just how reflective control works—just how is it that reflecting upon the justification of one of your attitudes

[27] Absent the stipulation of rationality, reflective control also seems to involve the familiar possibility of failure. With that stipulation, it displays the kind of invincibility sometimes thought to be distinctive of autonomous agency.

can change the attitude? Understanding this proves more difficult than is sometimes noted.

Sometimes people talk about the "authority of reflection" or "command of reason," as though a reflective judgment serves as an authoritative decree that one's attitudes obey, insofar as one is rational. But such talk is surely metaphorical. Retreating from the metaphor, people sometimes simply say that, when we reflect and find that some attitude of ours is unjustified, we then "correct" or "revise" or "update" the attitude under reflection. But the question at issue is, just how do we accomplish this correction or revision? What sort of activity or agency is exercised in such correction or revision? As already noted, the correction would not be well modeled as an action of the ordinary sort—we do not find ourselves with a bad attitude and then decide to change it by performing some action, as though we were changing a bad spark-plug. Exercising reflective control over one's own mind is not like surveying and tinkering under one's own hood.[28] It is not, to drop the metaphor, an exercise of managerial control—of taking action so as to affect one's mind.

The difficulty here should not be underestimated: the problem with modeling the correction or revision of one's attitudes as an action is not that it is hard to see what sort of *process* the corrective or revisionary action would involve; the problem is not alleviated by, e.g., thinking that the action of correcting one's own attitude is a basic one, which can be accomplished simply by deciding to do it. (Attempting to alleviate the problem in this way will return one to the metaphor of command: one will think that correcting one's attitudes is, after all, like surveying and tinkering under your own hood,

[28] This point was suggested to me long ago by Richard Moran.

so long as you are endowed with godlike powers to effect the required changes just by deciding that it be so.) Rather, the problem is that the correction or revision of an attitude is not well modeled as an intentional action at all, whether basic or complex. First, correcting or revising your own attitudes is not voluntary (in contrast, e.g., to correcting or revising your own speech): you cannot decide whether to correct or revise for any reason you think shows it good to do or not do. Perhaps, e.g., you think there is very good reason to leave some error in place. It does not seem open to you to do so, given the stipulation of rationality. Further, if the correction or revision were an intentional action, it would be accomplished by settling for oneself the question of whether to correct or revise. The correction, then, seems to be initiated by an exercise of evaluative control—but we were appealing to reflective control precisely to try to understand the most fundamental exercise of mental agency.

A somewhat more promising route employs the thought that, if you find that an attitude of yours—a belief that *p*, say—is unjustified, you will therein form a second-order attitude about that belief: a belief that your belief that *p* is unjustified. (I would say that you form this second-order attitude by an exercise of evaluative control.) One might think that, given this higher-order thought, simple compliance with the requirements of rationality will ensure that you do not go on believing that *p*. Insofar as one is rational, we might say, the lower-order attitude is "sensitive to" the higher-order judgment.[29] Thus, it might seem, once a rational creature is capable of reflection, it gains a kind of

[29] A powerful presentation of this thought, using the notion of "judgment-sensitive attitudes" can be found in T. M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), especially chapter one.

control over its own mind: insofar as it is rational, its mind will conform to its own reflective thoughts about how it should be.[30]

This is a powerfully attractive account. Again, I believe it is powerfully attractive in large part because it preserves the familiar features of the paradigmatic exercises of agency. It is easy to imagine that the agent is the one reflecting, making the higher-order judgment, and it is easy to think of that higher-order judgment as in some way affecting the lower-order attitude, so long as one is rational. Making a judgment and thereby effectively changing an attitude seems a lot like acting upon an object or issuing an effective command, either of which are obvious exercises of agency.

Though this is a powerfully attractive account, it should be noted that the question at hand—just how *does* one correct or revise one's attitudes under reflection—is not clearly answered by it. The account simply stipulates that one has satisfied the standards of rationality, and notes that these standards require a change in attitude. But we were wanting to understand how it is that one changes that attitude, given that doing so is not a matter of performing a kind of mental action. So, even if we grant (what, below, I will suggest we should not) that, if one is rational, one is sure to change one's first-order attitudes upon making the higher-order judgment, we will not thereby have come to understand the agency by which we conform to that requirement. This lack is made more worrisome by the fact that the simple fact that one's mind is functioning in accordance with certain standards does not typically show that one has exercised agency. (The well-

[30] The ability to have the mind one wants is the cornerstone of what has been a very fruitful line of thought over the last several decades. See, e.g., Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no. 1 (1971) and Charles Taylor, "Responsibility for Self," in *The Identities of Persons*, ed. Amélia O. Rorty (Berkeley: University of California Press, 1976).

functioning of one's perception or one's memory, e.g., does not seem to be, itself, an exercise of agency.)[31]

One might reply that the the functioning of a mind in accordance with the standards of rationality *just is* the activity of an agent. While I have some sympathy with this thought, I worry about its implications for our moments of irrationality. Still, even if we were to grant this thought, we might raise others worry about the picture at hand:

First, in order to preserve the familiar features of ordinary agency, rationality must in some way privilege the higher-order, reflective thought over the lower-order, unreflective one. But it is not clear why the requirements of rationality should display such a bias. Perhaps it is your lower-order thought that is rational and reasonable, and your higher-order one that is paranoid, compulsive, or self-deceived. In such a case it seems that the requirements of rationality might ask you to persist in your lower-order thought and abandon the higher-order one.

Additionally, and familiarly, it is unclear why sensitivity to a higher-order thought should render a lower-order attitude the product of one's agency or control unless the higher-order thought is itself already an instance, embodiment, or product of agency, or unless the agent is in some way already identified with it, such that its effects can be identified as hers.[32] Taking the first route—explaining why the higher-order thought is

---

[31] One might argue that the revision of the lower-order attitude is an exercise of agency in the following way: The well functioning of one's mind, one might say, ensures that that exercise of agency has its natural effects, in much the way that the well functioning of one's musculature ensures that one's intentions have their natural effects. Thus, just as actions are exercises of agency, so is the revision of the attitude.

This line of thought overlooks an important difference between the action and the revision of the first-order attitude, a difference which appears when things do not function well. If one's musculature fails in some way—seizes or spasms—one's action fails, but one is in no way irrational. But if one fails to revise one's lower-order attitudes, one is irrational.

[32] This is structurally similar to the point made by Gary Watson against Frankfurt's hierarchical account of free action. See Gary Watson, "Free Agency," in *Free Will*, ed. Gary Watson, *Oxford Readings in*

itself already an instance, embodiment, or product of agency—will, I think, lead one back to notion of evaluative control (or something very much like it).[33] The second route—identifying the agent with the higher-order attitude, such that its effects are hers—has been taken by a number of people.[34] Notably, those who take this route appeal, not to reflective judgments, but rather to desires or values. Moreover, I believe they do so advisedly: values and certain desires can plausibly be claimed to be the sort of thing with which an agent is essentially identified—something an agent cannot coherently disavow. A higher-order judgment about the justification of some other attitude, in contrast, does not seem to be the sort of thing with which an agent is essentially identified; it seems, rather, like something one can coherently disavow.[35]

Rather than trying to further explain and defend my dissatisfaction with going accounts of reflective control, I will, at this point, simply present the beginnings of an alternative. I believe the account of evaluative and managerial control that I have offered above might provide a particularly promising way to start to understand reflective control. Insofar as this alternative is plausible, it raises another worry for what I have called the powerful picture.

Philosophy (Oxford: Oxford University Press, 1982). Frankfurt and his followers retrenched by attempting to identify the agent with some (set of) attitude(s), which then relate to others. Frankfurt's later work is found in his *The Importance of What We Care About* and in his *Necessity, Volition and Love* (Cambridge: Cambridge University Press, 1999). J. David Velleman develops the thought in a different way. See *Practical Reflection* (Princeton: Princeton University Press, 1989) and *The Possibility of Practical Reason*.

[33] Recall that the appeal to reflection was an attempt to preserve the standard features. We have now found, in attempting to work out the reflective account, that we need to appeal to an exercise of agency within it—agency in forming a judgment. If we try to secure the standard features here, we risk generating a regress.

[34] Central examples are, again, Velleman and Frankfurt.

[35] My treatment of such alternatives here is obviously only provisional. I hope to give them a fuller hearing in later work.

Consider, then, successfully revising one's belief that *p* under reflection. One believes *p*, we have said, just in case one is committed to a positive answer to the question of whether *p*. To conclude that one's belief that *p* is unjustified is to conclude that one does not have sufficient reason to settle that question positively. Thus, to conclude that one's belief that *p* is unjustified is to settle negatively the question of *whether the reasons available to me show that p*. But it seems that settling this question might involve reconsidering the simpler question, *whether p*, while employing the reasons available to you. That is to say, reaching the conclusion that your belief that *p* is unjustified might itself involve reconsidering the basic question, whether *p*, and failing to settle it positively. If it does, then, insofar as you remain consistent, or of one mind, on the root question of whether *p*, you will, in failing to settle the question positively, therein suspend your belief that *p*. That is to say, insofar as you remain of one mind on the question of whether *p*, you might revise your belief *in the process* of finding it unjustified.[36]

[36] Yannig Luthra and Sheldon R. Smith independently suggested that a person might conclude that she does not have sufficient reason to believe *p* without re-posing the question of whether *p*: Perhaps she now decides that any belief she acquired last night, when exhausted and under the influence of all those pain-killers, must be unjustified. And suppose she knows that last night she acquired the belief that she will recover fully from her accident in two weeks time. She might now conclude that she does not have sufficient reason to believe that she will recover fully in two weeks time, and thereby lose that belief. But it might seem that she has revised her belief without re-posing for herself the question of whether she will recover fully in two weeks—she answered for herself a question about the justification of her belief, without reconsidering the truth of the matter (as Yannig nicely put it, she has reasoned as a juror, about the adequacy of her evidence, not a detective, about the facts)—and so this might seem a different kind of case than the one I consider in the text. If so, then in such a case the question I have been asking remains unanswered: how, exactly, does the person revise her unjustified belief? What kind of agency is at work, in the revision?

But I am not sure that, in drawing the conclusion that, because of all those pain-killers, she does not have reason to believe she will recover, our patient does not thereby reconsider whether she will recover. (I am not sure that, with respect to her own beliefs, she can reason only as a juror, and not also as a detective.) In any case, because I am also inclined toward the stronger thesis mentioned in footnote X, I am inclined to think that whenever a person revises her belief that *p* the person will have exercised evaluative control over her belief that *p* (because, in revising her belief that *p*, the person must have revised her commitments about whether *p*, and, if the stronger these is true, the revision of such commitments is accomplished by an

Understanding reflective control along these lines has several benefits. We have accounted for the change in the first-order attitude by appeal to an exercise of evaluative control together, not with the requirements of rationality (quite generally), but rather with a (weaker) requirement of consistency. We stipulated, not that rationality privileges the higher-order judgment over the lower-order attitude, but simply that the person stays of one opinion on the root question, as he or she settles the more sophisticated question. Further, because it is clear that evaluative control is being exercised as the person addresses the sophisticated question, there is no need to identify the agent especially with the higher-order judgment. We rather simply identify the agent as the one exercising evaluative control.

The proposed account also goes somewhat further than the alternatives in answering the question with which we started: just what form of agency is exercised over the attitudes revised under reflection? The proposed account would have it that one exercises evaluative control in revising one's attitudes under reflection.

Note that, insofar as the alternative picture is close to correct, the original models of reflective control are not just metaphorical, but actually misleading. We started by appeal to the metaphor of commanding or tinkering. But notice that any commanding of or tinkering with attitudes must be *subsequent to* the judgment that the attitude is unjustified: one first makes the judgment and then commands or acts upon the attitude that one has judged unjustified. The unjustified attitude appears, in these metaphors, as an ordinary object of manipulative control. But, on the proposed account, the revision of

exercise of evaluative control). If so, then she will have reconsidered the question of whether *p*. But since I am not prepared to advance the stronger thesis, I will leave it that the agency exercised in revising the belief that *p*, in the case imagined, might remain a bit of a mystery. I will be relatively happy if I have provided a clearer account of at least one way in which reflective control is exercised.

the belief is not subsequent to the making of the judgment; it is accomplished in arriving at that judgment. Once the judgment has been formed, there is nothing left to command nor anything with which to tinker.

Likewise, if the alternative picture is correct, the powerful picture according to which one's lower-order attitudes conform or are sensitive to one's higher-order judgments (insofar as one is rational) is misleading in the same way. First-order attitude could properly be thought of as *sensitive to* the higher-order judgment, because, insofar as one remains of one mind, the first-order attitude will be revised or suspended in the process of arriving at the higher-order judgment. When things go well, the attitude and the judgment do not cohabit the mind. At best, the attitude is sensitive to a stretch of the reasoning that supports or generates the higher-order judgment.

Conclusion

Given the scope of the topic, my aims have been modest. I hope to have introduced a way of thinking about our agency over certain of our attitudes that I have found fruitful. This way of thinking requires a certain assumption: the assumption that certain of our attitudes embody our answer to a question of set of questions. Given this assumption, it seems we will exercise agency over these attitudes in two distinct ways: by changing our answer to the question(s) they embody or by acting upon them so as to affect them according to our purposes, in roughly the way we can act upon any object that interacts in more-or-less predictable ways with its environment. The first I call exercising evaluative control over the attitude; the second I call exercising managerial or manipulative control.

These two forms of agency are rarely distinguished, because evaluative control does not display the most familiar features of agency while managerial or manipulative control

seems to involve an exercise of evaluative control (perhaps more than one). I hope I have suggested why evaluative control deserves to be thought of as a form of agency, despite the fact that it does not sport the usual features. I have also tried to make clear how exercises of managerial control can involve an exercise of evaluative control. Finally, I hope I have shown how certain complex exercises of agency over our minds, including what I have called reflective control, might be modeled in terms of these somewhat simpler forms of agency.[37]

REFERENCES

Albritton, Rogers. "Freedom of the Will and Freedom of Action." In *Free Will*, edited by Gary Watson, 408–23. Oxford: Oxford University Press, 2003.

Anscombe, G. E. M. *Intention*. Oxford: Blackwell Publishing Co., 1957.

Boyle, Matthew. "Making up Your Mind." (in process).

Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge: Cambridge University Press, 1987.

Feldman, Richard. "The Ethics of Belief." *Philosophy and Phenomenological Research* 60 (2000): 667-96.

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no. 1 (1971): 5-20.

———. *The Importance of What We Care About*. Cambridge: Cambridge University Press, 1988.

Frankfurt, Harry G. *Necessity, Volition and Love*. Cambridge: Cambridge University Press, 1999.

Hieronymi, Pamela. "Controlling Attitudes." *Pacific Philosophical Quarterly* 87, no. 1 (2006): 45–74.

———. "Responsibility for Believing." *Synthese* 161, no. 3 (2008): 357–73.

Kavka, Gregory. "The Toxin Puzzle." *Analysis* 43 (1983): 33-36.

Moran, Richard. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press, 2001.

Owens, David. *Reason without Freedom: The Problem of Epistemic Normativity*. London: Routledge, 2000.

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

Shah, Nishi. "How Truth Governs Belief." *The Philosophical Review* 112 (2003): 447–82.

Taylor, Charles. "Responsibility for Self." In *The Identities of Persons*, edited by Amélia O. Rorty, 281-99. Berkeley: University of California Press, 1976.

[37] This paper has benefited from the helpful comments and questions of many, including Michael Bratman, Denis Bühler, Tyler Burge, Stephen Darwall, Sean Kelsey, Niko Kolodny, Yannig Luthra, Sheldon R. Smith, the Philosophy Department at UNC Chapel Hill, the participants of the 2007 SPAWN conference, and an anonymous reviewer at Oxford University Press.

Velleman, J. David. *Practical Reflection*. Princeton: Princeton University Press, 1989.

———. *The Possibility of Practical Reason*. Oxford: Oxford University Press, 2000.

———. "What Happens When Someone Acts?" In *Perspectives on Moral Responsibility*, edited by John Martin Fischer and Mark Ravizza, 188-210. Ithaca: Cornell University Press, 1993.

Watson, Gary. "Free Agency." In *Free Will*, edited by Gary Watson, 96-110. Oxford: Oxford University Press, 1982.