Philosophical Perspectives, 14, Action and Freedom, 2000

FREE WILL AND AGENCY AT ITS BEST

Gideon Yaffe
University of Southern California

Some kind of freedom—call it "freedom of action"—is undermined by ropes, chains and other physical constraints. And there is something that all of these constraints have in common—something that makes them all *constraints*: they stand as obstacles to the realization of certain choices.[1] Freedom of action, then, is dependency of conduct on choice; the reason that physical constraints undermine this kind of freedom is that they interfere with such dependency. The question needing to be tackled is this: What more does a full-fledged free agent—an agent who has all the kinds of freedom that we worry about when we worry about "the free will problem"—need beyond this rather limited kind of freedom? Sometimes nothing but fear and apathy, for instance, prevent us from coming to another's rescue; sometimes we are less than full-fledged free agents despite the fact that we are not tied down, our phones lines are not cut. We can be unfree even when nothing interferes with the efficacy of our choices, for things like fear and apathy can perniciously influence what we choose. What we lack in such circumstances is freedom *of will*. But what, if anything, is freedom of will?

This paper argues that two broad strategies for answering this question are mistaken. Many philosophers who have offered substantive theories of freedom of will have followed one or the other of these two strategies. The result is that we need a new approach. Towards the end, I suggest an alternative approach and give some reasons for thinking it promising.

The strategies which I am attacking here, and that which I propose, all depend on a particular conception of what we are doing when we offer an analysis of freedom. In particular, I assume throughout that a choice or action possesses a particular kind of freedom because of something about either what causes it, or the manner through which it is caused. We know, roughly, what feature an action's causal history must exhibit to be an instance of freedom of action: it must be caused by a choice to do it, where the choice is causally crucial: in the absence of the choice the agent would not have so acted. But what must the causal history of a choice be like for the choice to manifest freedom of will?

What this approach implies is that those with a certain kind of incompatibilist bent are simply not going to be convinced by the view which I propose any more than they are convinced by the views which I attack. That is, there are those who think that the question of an agent's freedom with respect to what she does or chooses turns entirely on the modal limits placed on her potential choice or action by her circumstances or psychology. They see the question of whether or not an agent is free as being the question of whether or not, given relevantly similar circumstances and a relevantly similar psychological state, she "could" have chosen or done otherwise. For theorists of this sort, the central question about freedom is really a modal question turning on the precise modality expressed by the word "could". There are, of course, many senses of "could" under which it is always false that we "could" have done or chosen otherwise than we did given the very same circumstances—if, for instance, the only things that "could" occur are those that did occur—and there are other sense in which it is sometimes true and sometimes false depending on the nature of our circumstances and psychological states. The discussion here simply doesn't speak to those who construe the free will question as, ultimately, a question about the precise modality expressed by the word "could".

Answers to broadly metaphysical questions—such as the question of what it is to possess freedom of will—are often driven by some intuitive, but unsystematic, prior conception of the nature of that which is to be explained. Accounts of the nature of freedom of will have tended to fall into one of two camps, corresponding to two different pictures of what freedom of will is. For some, freedom of will is to be equated with self-expression in choice. According to this picture, the more our choices are *ours*, are grounded in and arise from something important about us, the more we approach freedom of will. For others, freedom of will is to be equated with self-transcendence. The agent who has freedom of will, on this conception, has a will that is responsive to and aimed at those aspects of her circumstances that are of genuine value, that are worthwhile guides of her choice; she is not a slave to herself, but manages to allow her will to express that which is worth expressing.[2]

Self-expression views are aimed at providing accounts of self-*determination* that are consistent with roughly naturalistic metaphysical assumptions. For self-expression theorists, self-determination is to be understood as causation of choice by certain events and states which, because of their relations to one another or to other states and events of the agent, constitute the kind of self that is crucial for moral responsibility, or for other forms of assessment that we might want to levy on an agent in response to a judgment of her freedom. According to such views, it is because her actions or choices express—that is, depend causally upon—deep structures in, or even constitutive of, the self that the agent can be said to have freedom of will.

I use the term "transcendence" to refer to self-transcendence views because for those who believe freedom of will to be equated with self-transcendence—with responsiveness to the evaluative or appropriately reason-

giving facts—the ultimate explanation for the claim that a particular agent's choice manifests freedom of will appeals to the fact that her choice is pegged to some fact not about her, but about her environment. Susan Wolf gives an example which might help here.[3] Wolf describes two different agents both of whom answer "Carson City" to the question "What is the capital of Nevada?". The first gives this answer because she has been taught to so answer, the second because, in fact, the capital of Nevada is Carson City. The explanation for the second agent's answer appeals to the truth of the answer; the fact that she gave the answer she did is explained by citing the fact that that is the right answer. Similarly, the choices of the self-transcendent agent can be explained by citing the fact that what she chose was evaluatively optimal; she chooses as she does because what she chooses is best. Since egoistic subjectivism about value is false—what is best for an individual to choose is never a function solely of facts about herself—the choices of the self-transcendent agent are thought to be free not because they arise out of and depend upon aspects of herself (although they may) but, rather, because they arise out of and depend upon those features of her circumstances and psychology on which the value of what she chooses supervenes.

These two traditions of thought on the nature of freedom of will point to two distinct strategies for offering a philosophical analysis of the concept. Those who follow the first, hold self-expression to be both necessary and sufficient for freedom; those who follow the second, hold self-transcendence to be. Substantive theories of freedom of will can be developed by following these strategies and providing criteria which must be satisfied by self-expressive and self-transcendent agents, respectively. But, as I argue here, no matter how these criteria are formulated, both strategies are fundamentally flawed. Neither self-expression nor self-transcendence is either necessary or sufficient for freedom of will. It would be a mistake, however, to simply abandon these strategies entirely, for self-expression and self-transcendence are, somehow, relevant to an analysis of the concept of freedom. I go on to suggest that what this shows, although not definitively, is that a particular assumption about the nature of the concept of freedom of will is false. The debate over the nature of freedom of will has proceeded from the assumption—sometimes explicit sometimes implicit—that freedom of will is a descriptive concept, a concept of metaphysics. But, perhaps this is false. Perhaps, rather, freedom of will is a "thick concept"[4], a concept that is not purely descriptive, but also imputes a certain form of value to that which falls properly under it. To reach this conclusion, I claim, is to draw an inference to the best explanation: there are certain facts which are best explained if freedom of will is a "thick" concept. But more of this later.

My discussion is informed by a distinction between desire and will. On one natural conception of the will—although not the conception that I will be using here—all desires or desiderative attitudes are acts of will. This is to countenance a distinction between beliefs—attitudes with mind-to-world direction

of fit—and desires—attitudes with world-to-mind direction of fit[5]—while denying there to be any important difference between willings and desires. On this conception, to want something is to will it. For my purposes here, however, it is important to note subcategories among the generic category of mental states with world-to-mind direction of fit. "Desires", as I will be using the term, are occurrent mental states that motivate us to act in ways useful (relative to our beliefs) for their satisfaction, but which are not governed by the same norms of consistency and coherence that govern willings.[6] It is irrational, for instance, to will both A and not-A, while it is not irrational to desire a state of affairs and at the same time desire that it not occur[7]. While there may be certain norms of rationality that govern desires, they are not the norms of consistency and coherence that govern choices. So, I divide the class of motivational states as follows: All motivational states have a world-to-mind direction of fit. Among those, some are governed by certain standards of rationality such as coherence and consistency and some are not. Those that are are willings, and this class includes a fairly wide range of mental states—choices, volitions, intentions—that may differ from one another but not in ways that are important for our purposes. Those that are not are to be called "desires".

It follows that it is possible not to will, or even to will contrary to, an action that one performs. While we may never act without a motive, we may very well be motivated by a desire rather than a choice, volition or intention. Notice that the word "action" is being used here in a broad sense. Volitional theories of action, for instance, reserve the word "action" for those events or states that are caused by volitions or other willings. There is much that is right about volitional theories, but, I will use the term "action" to refer to any state or event caused appropriately by a motive (although I simply won't give flesh to the word "appropriately" in this formulation) and leave open the possibility that states or events caused appropriately by willings have some important status (perhaps such events are *intentional* or *voluntary* actions).

One final preliminary point. In a number of places I draw on the idea of action performed "for a particular reason". I assume that the relation between a reason, or a mental representation of a reason, on the one hand, and an action performed for that reason, on the other, is causal. To act for a reason is for one's action to be caused in some special way by the feature of the world that is reason-giving, or by one's mental representation of a reason-giving feature, or, perhaps, by both. This is not a substantive theory of what it is to act for a reason, for I have not said what the relevant causal relation is. What is important for my purposes is only that actions performed for reasons have no special features that cannot be countenanced by a fully naturalistic theory of the mental. Just as an action's or choice's freedom is to be found in features of it's causal history, so too the distinctive mark of acting for a reason is to be found in features of the causal relation between the reason (or mental representation of it) and the action performed for the reason.

A Lesson from Standard Compatibilism

In the history of philosophy, at least one figure of towering importance thought that we didn't need a theory of freedom *of will* at all. Hobbes held that freedom of action was both necessary and sufficient for full-fledged freedom.[8] Hobbes was probably wrong. One way of diagnosing his error helps us to see the mistake that is made in equating self-expression with freedom of will. Let's formalize this view as follows:

> *The Standard Compatibilist Thesis*: An agent has full-fledged freedom with respect to action A iff (1) If she chooses to A she will A, and (2) If she does not choose to A, she will not.[9,10]

Even Hobbes encountered the following objection to this view[11]: the truth of conditionals (1) and (2) is not sufficient for freedom. There are unfree agents made unfree not by things like ropes and chains, but, instead, by psychological disorders, coercion, indoctrination, sometimes childhood trauma, and, in certain circumstances, even ignorance of the facts. But these forces undermine freedom not because they are obstacles to the realization of choices, but because they perniciously influence choice. Standard Compatibilism ignores, or dogmatically denies, that this is even possible.[12]

The Standard Compatibilist can respond to this objection by claiming that even the pernicious factors (psychological disorders, coercion, etc.) are obstacles to the realization of our choices and, hence, make the relevant conditionals false. And, this response seems, at first glance, to be satisfactory if, to take one illustrative example, coercive threats lead to action without that action being chosen. Perhaps, the Standard Compatibilist might say, when someone receives a coercive threat, overpowering desires are raised in her which bring her to comply with the threat regardless of what she chooses to do. Her desires control her action, the Standard Compatibilist might insist, thus detaching her will from the causal etiology of her compliant performance. Or, while she may choose to comply with the threat-induced desire, the desire would have, itself, been sufficient to bring about her compliant action regardless of what she chose. This Standard Compatibilist response relies on a seemingly implausible account of coercion, but what, exactly, is wrong with it?

One way to respond to this question is to begin with another: Is coerced action, on the Standard Compatibilist's analysis, intentional action? If not, then, on the Standard Compatibilist's analysis of coercion, coercion undermines freedom by undermining the intentionality, or voluntariness, of the actions it induces. The trouble is that many coerced acts that are rightly described as unfree—acts that are excused with expressions such as "I had no choice, he had a gun to my head"—are intentional. An agent might coolly calculate the results of non-compliance and decide, in the end, that it is better to do as the

coercor demands, and, nonetheless, act unfreely. If the force that the coercor applies is sufficient to insure compliant action then the agent is unfree even if the coercor insures compliant *intentional* action. To utilize this response to the Standard Compatibilist is to shift the ground of debate from a question about freedom to a question about intentional action. It remains theoretically open to the Standard Compatibilist to insist that it is possible to act intentionally without one's will playing a crucial role in the causal etiology of action. Thus, the Standard Compatibilist might say, coerced actions are intentional, yet unfree, precisely because they are intentional actions brought about without the participation of the will of the agent. Is there something wrong with the Standard Compatibilist's analysis of coercion even if we grant her the possibility that coerced action, so analyzed, can be intentional action? That is, can we respond to the Standard Compatibilist's analysis of coerced action without shifting the debate to consideration of the nature of intentional action? The answer is "yes".

Start by distinguishing among three things: (1) the behaviors on the part of a manipulator that make it right to say that that individual is engaging in coercion of another, (2) the action that the manipulated performs as a result of and because of the pressures applied by the coercor, and (3) the causal mechanism through which the coercor's behavior succeeds in inducing the compliant actions of the coerced. The Standard Compatibilist provides an explanation for the unfreedom of (2) by appeal to the features of (3); that is, the Standard Compatibilist explains the unfreedom of the coerced conduct not by appeal to the fact that the conduct is coerced but rather by appeal to *the means through which* the coercive pressures are claimed to have their effect: they induce over-powering desires. However, imagine that an agent acts to comply with a coercive threat for the reason that the coercor has provided: I hand over the money precisely so as to avoid being shot. To know that an agent acted for the reasons supplied by the coercive manipulator is to know that the behaviors referred to in (1) cause (2) in some distinctive (undisclosed) way; recall, after all, that that is all there is to acting for a reason. But to know that I acted unfreely in response to coercion it is sufficient to know that the coercive behaviors provided my reason for acting as I did. In order to determine that I am unfree, we don't need to know (3); that is, we don't need to know how, precisely, it came to pass that I complied with the manipulator's demands in order to know that I am unfree. Thus, the Standard Compatibilist offers an explanation for the freedom-undermining force of coercion which is not consistent with an important pattern in our flow of concepts: our move from "acted for the reasons supplied by the manipulator" to "acted unfreely" is not mediated by any further concept such as "acted as a result of over-powering desires".

Notice that it does matter to freedom how coercion causes compliant action. That is, the coercive behaviors must provide the reason for the compliant action, and must, therefore, cause the action in some distinctive way. However, the features of the causal route from coercive pressure to compliant action that the Standard Compatibilist appeals to are not the distinctive features of the route

from a reason to action performed for that reason. We often act for a particular reason without being motivated by over-powering desires. The point is that the freedom-undermining force of coercion cannot be explained by appeal to any features of the causal chain from coercive pressure to compliant action not already possessed by that causal route merely by virtue of the fact that it constitutes action for the reasons supplied by the coercion. What this implies is that, likely, the feature of the causal history of coerced acts by virtue of which those acts are unfree will be found merely in the fact that *coercion* is in their causal history and not in the facts about the particular way in which the coercion causes them. So long as it is true that it is sufficient for unfreedom to act for the reasons supplied by the coercor, this must be true, because there is nothing about acting for a reason generally which undermines freedom.

In the discussion so far, much is being built into the idea that the manipulator *supplies* the reasons for the action. We often "supply"—in some sense of "supply"—the reasons for the choices of another without thereby coercing that other: the chef supplies my reasons for choosing to buy the food by cooking it so beautifully; my wife supplies my reasons for choosing to take a trip by being the one that I want to travel with. Further, it is very difficult to specify the precise difference between the way in which a coercor "supplies" reasons for action and the way in which individuals supply such reasons in these cases. But such a specification will be part of an account of the nature of coercion. To coerce someone is to effectively supply her with reasons to act in some special sense of "supply" that I am not specifying here. But when someone else supplies one's reasons for action in the way that is distinctive of coercion, and one chooses in accordance with the reasons so supplied, one chooses without freedom of will. What this means is that the unfreedom which agents experience as a result of coercion comes from the fact that coercion is the source of one's reasons and not from features of the causal route through which those reasons induce compliant choice.

The Standard Compatibilist, then, provides a sufficient condition for the freedom-undermining force of coercion, but doesn't supply the *important* sufficient condition, for her explanation doesn't explain why, in general, coercion undermines freedom whenever the agent acts for the reasons supplied by the coercor's pressures. What this suggests is a very general methodological principle: When providing an explanation for why X undermines freedom, whenever an agent who acts for the reasons that X supplies acts unfreely, we must find a connection between X itself and unfreedom and not a connection merely between regularly found features of the causal sequence from X to action and unfreedom.

So Standard Compatibilism fails, but fails constructively, for its failure points to the need for care when considering how one's theory of full-fledged freedom accounts for the freedom-undermining influence of some particular feature of oneself or one's environment. This methodological lesson will become important in the next section.

Freedom of Will I: The Conditions of Self-Expression

One approach to understanding the nature of freedom of will starts with reflection on what is attractive about the account of freedom of action sketched already. While that account does give an explanation for the fact that ropes and chains undermine freedom, we can still look deeper: why do we take dependence of conduct on the will to constitute *any* form of freedom (even if not all that we take full-fledged freedom to be)? One possible answer—and there is another possible answer to be discussed in the next section—is this: when an agent's conduct depends on her will then part of what happens in the world tracks something about her; to know that an agent had freedom of action when she acted is to see particular occurrences (her bodily movements and certain of their results) as expressive of something about her: what she willed. On the flip side, to know that she had freedom of action with respect to an action that she did not perform is to know that something that was possible failed to happen because of something about her. When we have freedom of action, the way of the world (or at least some of it) is expressive of the state of our wills. The agent who has freedom of action, then, expresses an aspect of herself.

Encouraged by this result, we might turn to the cases of psychological disorder and the rest to see whether, perhaps, what those forces undermine is some other, deeper form of self-expression. And, in fact, it is something like this project that is undertaken by Harry Frankfurt in his widely read paper "Freedom of the Will and the Concept of a Person".[13] It is not enough for an agent's choice—what Frankfurt analyzes as, merely, "effective first order desire"—to be expressed in her conduct for her to be free; deeper facts about her—facts about her reflective attitudes—must be expressed in her choices. Her choices must arise from and depend upon[14] these deep structures in the self, thinks Frankfurt, if she is to have freedom of will. In later work[15], Frankfurt has made further efforts to specify both what structures of the agent need to be expressed in her choices and what relationship her choices must bear to those structures if the agent is to have freedom of will. And similar efforts have been exerted by other theorists as well.[16] But, as I argue in this section, it is neither necessary nor sufficient for the possession of freedom of will to be an agent whose choices are self-expressive.

To see that self-expression is not sufficient for freedom of will, think of the very cases that impugned the case for Standard Compatibilism: the cases of psychological disorder, coercion and indoctrination. An agent who is subjected to certain forms of brainwashing may come out of the treatment a fully integrated and wholly devoted subject, willing, perhaps, to sacrifice all in order to protect any hair on the head of her Leader. When such an agent goes through with it—whatever it is, and it could be *anything*, that is just what makes the very idea of brainwashing so nightmarish—she lacks some kind of freedom...at least, so it seems. It is not freedom of action that she lacks (we can suppose), so she lacks freedom of will. But why? Those who wish to analyze freedom of

will as consisting in self-expression in choice must say one of two things in response: (1) the brainwashed don't lack freedom of will (and, hence, since they do not lack freedom of action, they are full-fledged free agents), or (2) the choices of the brainwashed are not self-expressive, or not self-expressive in the right way, or to the degree that they need to be.

Those who take the first route—who deny that the brainwashed lack freedom of will—are denying that it is even possible for someone who possesses the right kind of psyche—who enjoys the specified relationship between deep psychic structures and choices—to lack freedom of will. They hold that it is inconceivable for someone to purposefully induce those structures and relationships with the intention of making the agent choose in a certain way as a result of and in accordance with those structures *and* thereby undermine her freedom of will. We can see what is wrong with this position with a thought-experiment. Imagine that you are a self-expression theorist and you take it to be both necessary and sufficient for freedom of will that an agent's choice bear relation R to psychic structure C. Now imagine that you are given information about a particular case in stages. Stage 1: At the moment, agent S has no inclination or desire to choose to A, nor ought she according to any fair normative standard. Stage 2: Cruella has targeted S and decides to do what it takes to get S to choose to A. Stage 3. Cruella develops a plan of attack and executes it. Stage 4: S chooses to A as a result of Cruella's machinations. Stage 5: Cruella got S to choose to A by causing S to exhibit feature C, which, given the circumstances, was sufficient for S to choose to A and for her choice to bear R to C.

Given the information in Stages 1–4, it seems appropriate to view S as a pawn in Cruella's hands: Cruella aims at nothing but S's making of a choice that S has no reason, prior to Cruella's machinations, to make. And Cruella gets her way. How could this appear to be an instance of freedom of will on S's part? And if it does not, then how could it help to be told *how* Cruella pulled it off, as you are told in Stage 5? Notice that there are various elements of the case as described that might turn out to be important. Perhaps it is important that Cruella *aims* at inducing compliance of the relevant sort and succeeds. That is, perhaps Cruella would not undermine S's freedom of will if she, say, performed actions not intended to induce S's compliance but which happened, nonetheless, to do so. Those who deny that it is possible to undermine an agent's freedom of will through brainwashing that induces the appropriate "mesh" between deep psychic structures and choices must deny that factors such as these are relevant: a choice that comes about as a result of the right kind of psyche exemplifies freedom of will no matter what the source of the relevant psychological configuration.

The trouble with the second possible response—the assertion that those who are the victims of brainwashing are not actually expressing themselves appropriately—can be seen by asking the following question: What aspect of such an agent is failing to be expressed in her compliant choice? Whatever answer is offered—her genuinely reflective acts of identification, her true char-

acter, her considered self-conception—those aspects of the agent can, also, be under the control of the brainwasher; if they are, then they too might be expressed in the compliant choice of the agent, for exactly what the brainwasher does is to cause the crucial elements of the agent's psyche to be pointed towards compliant action. We might put this point in terms of a challenge: What is the difference between a fully-integrated and unconflicted agent who has become so through some neutral process of training—an agent, for instance, who has the Aristotelian virtue of bravery—and an agent who is just as unconflicted and devoted to a certain course of conduct as a result of brainwashing? We want to say that the former has freedom of will and the latter does not, but how can we say that within a theory that sees freedom of will as consisting in self-expression in choice?

We might try to answer this challenge by insisting that there is some crucial psychological difference between these two types of agent, a difference which manifests itself as a difference in self-expression. We might say, for instance, that the choices of the brave are available to revision under reflective scrutiny in a way that the choices of the brainwashed are not. Hence, the choices of the brave—in contrast to those of the brainwashed—express not just what the brave are in fact devoted to, but also what they would be devoted to were they to reflect in various ways. The trouble with this response is that it involves violation of the methodological principle described in the previous section. I explain.

We are assuming, at this point, that when an agent makes a choice for the reasons supplied by a process of brainwashing, she lacks freedom of will. The self-expression theorist might explain this by suggesting that brainwashing causes choices that are not available to revision under self-scrutiny; even if the brainwashed were to recognize after self-scrutiny that the potential choice is faulty, she would still choose in that way. This explanation notes a particular feature of the causal sequence from brainwashing to compliant choice: it is rigid; it cannot be changed even as the result of critical examination that finds it to be faulty. But this is not a feature of every causal process from a reason (or a mental representation of a reason) to action for that reason; nor is it a feature which must be possessed by the causal sequence from manipulative behavior to compliant action in order for that manipulative behavior to count as brainwashing; and, yet, whenever agents choose for the reasons supplied by a process of brainwashing, they lack freedom of will. Thus, the explanation picks out a non-crucial fact about a very large set of cases of choice in response to brainwashing and claims it to be the crucial feature.

An example might help here. In the film *The Manchurian Candidate*, a man is brainwashed in such a way that, afterwards, whenever a phone rings in a certain manner, he chooses to do whatever it is that the voice on the other end tells him to do. Let's grant that the process is such that even if he were to reflectively examine the action he is told to do, and even if were to conclude it to be a horrible, unacceptable act, he would still choose to do it. But is this the reason that he is unfree? No. All we need to know in order to know that his

responsibility for his choice is severely diminished is to know that he chooses for the reasons supplied to him by the processes of brainwashing that he has undergone. If those are his reasons for choosing as he chose, then we know that he lacks freedom of will when he chooses, regardless of the features of the causal sequence from brainwashing to choice (except those that are required for him to be rightly said to be acting for those reasons).

The problem here is not specific to this particular explanation for the freedom-undermining nature of brainwashing, but is, rather, endemic to *any* explanation that the self-expression theorist might give. The problem is that the self-expression theorist tells us that when the causal chain leading to an agent's choice has feature Y, then that choice is self-expressive. Then, when faced with examples of causal chains leading to choice that have feature Y as a result of the manipulation of a brainwasher, the self-expression theorist insists that, in fact, brainwashing produces choices only through causal processes that *lack* feature Y, and, further, that is why the relevant choices are made without freedom of will. But, unless feature Y is incompatible with the special features of causal sequences by virtue of which those causal sequences are instances of taking as a reason, this response shows too much, for it shows that we aren't always unfree when our choices are made on reasons supplied by brainwashing, but only when they are supplied by brainwashing in a particular way. But this is false for reasons similar to those suggested by the Cruella example above: to know that a manipulator aimed at and succeeded in inducing a choice that the agent had no independent reason to make (the information described in stages 1–4) is to know that the agent's freedom of will is undermined even in the absence of knowledge of the particular means through which the manipulator succeeded in her endeavor.

We can see why self-expression is not even necessary for freedom of will by reflecting on the following case. "The Dutiful" is an agent who reliably chooses in the best possible way, all things considered, given her circumstances. Further, she not only chooses in accord with what is genuinely good, she tracks the good: for each of her possible actions, if that action were the best of her alternatives, she would choose it. The explanation for the fact that she made the choice that she made always appeals to the optimality of that choice. Often, although not always, the Dutiful chooses in a self-sacrificial way. Her particular desires do not take precedence over the desires of others; she chooses as she most desires only when what she most desires accords with what is genuinely best for her to do. Now imagine this person put to the test. Imagine, for instance, that, like Job, everything that she cares about, desires or hopes for is taken away from her, and all she needs to do to get it back is to, say, renounce God (assume this to be a non-optimal act). But she doesn't; she chooses instead to endure her suffering rather than to choose anything other than the best of her possible actions.

Is the choice of the Dutiful self-expressive in the requisite sense? What reason is there to think so? The Dutiful's choice is in conflict with every inclination she has: every desire, every whim, every hope that she has ever held or

entertained remains unexpressed, even thwarted, by her choice. We often make choices that fail to be expressive of some of our desires: I choose to work instead of going to the movies not because I don't want to go the movies—I do—but because I need to work. But the case of the Dutiful is different: in certain situations, the entire desiderative side of the Dutiful's psyche plays no role in the production of her choice and, hence, remains unexpressed, unrevealed, by what she chooses. Could an agent whose choices fail to express the entire desiderative side of her psyche be self-expressive? If not, then the case of the Dutiful shows that self-expression of the requisite sort is not necessary for freedom of will.

Again, there are two ways to resist this conclusion: (1) Deny that the Dutiful has freedom of will, or (2) Deny that the choices of the Dutiful fail to be self expressive in the requisite sense. Neither answer is satisfactory.

The trouble with the first of these responses is that there are patterns in our practices that favor the thought that the Dutiful possesses freedom of will. For instance, when faced with two agents each of which has endured the Dutiful's travails where one has chosen as the Dutiful does and the other has given in to the substantial pressures to renounce, whom do we pity and whom do we praise? It is the latter, and not the former, that seems to have been "swept away", or to have not been herself, and the former that seems to have stuck her feet to the ground and not allowed herself to be moved by the forces applied to her. While she labors under substantial pressures to choose otherwise than she does, she does not seem to be under any pressure to choose as she does. How can an agent be unfree when she acts contrary to all of the forces that seem to be pushing on her? It seems that the primary motivation to deny the freedom of will of the Dutiful comes from adherence to an equation between self-expression and freedom and not from examination of the facts.

The trouble with the second response—the insistence that the Dutiful is self-expressive—is as follows: What is it about the Dutiful which is expressed in her choice? Say that the answer to this question is "feature C"—perhaps C is a feature of the Dutiful's character, a deep-seeded evaluative belief, or a disposition to engage in deliberative activities that help her to recognize the good. What about the case forces us to believe that C is expressed by the Dutiful's choice? Well, it seems that the reason to believe C to be expressed is either that it is conceptually impossible to make a choice without expressing feature C, or else that it is conceptually impossible to choose *the good* without expressing feature C.

The former answer can't be right since it implies that it is not possible to choose without expressing oneself in the way that is taken to be necessary for freedom of will. But then this kind of self-expression is an idle condition on freedom of will: no one can lack freedom of will by failing to so express oneself in choice, since it is not possible to both make a choice and so fail. To advocate this response is to return to Standard Compatibilism.

The latter answer—one cannot choose *the good* without expressing feature C—is better, but it is still problematic. The easiest way to make sense of it

is to think of feature C as a capacity for recognition of the good together with the power to take the good as one's reason for action. It seems plausible enough that an agent could not really be said to be choosing the good if her choice was not expressive of capacities such as these. To express feature C, on this analysis, is to express something that any agent who chooses the good must express in order to be rightly said to be choosing the good. In a sense, feature C is given with a choice of the good. This suggests that the fact that the Dutiful's choice is self-expressive in this sense is not the crucial fact about the Dutiful that accounts for her freedom of will. The Dutiful is an agent who is self-transcendent. It is possible that an agent is only self-transcendent if in addition to expressing the evaluative facts in her choices, she expresses something about herself in her choices (feature C). But the question is this: which feature of the Dutiful is accounting for her freedom of will, her self-expression or her self-transcendence? If the latter, then self-expression is only necessary for the Dutiful's freedom of will because it is necessary for her self-transcendence.

Still, nothing said so far speaks definitively against this way of resisting the denial of the necessity of self-expression for freedom of will: so far, there is no reason to believe the Dutiful to be a counterexample to the claim that self-expression is necessary since she has freedom of will and is self-expressive in whatever way a self-transcendent agent must be in order to be self-transcendent. However, the trouble with this analysis of the case is that it doesn't allow the self-expression theorist to maintain a consistent account of the nature of the kind of self-expression thought to be necessary for freedom of will. To see this, consider another case: "The Aesthete" always chooses in such a way as to maximize aesthetic value in the world, and is utterly unaware of either moral or prudential value. The Aesthete, for instance, might live in abject poverty so as to fund an elaborate and promising project involving thousands of clones of Paul Cezanne. When asked questions about why she does this, she eloquently extols the virtues of Cezanne's work and complains bitterly of the tragedy of his death at the age of 67—"Ten more years, and who knows what wonders he might have produced!". When asked about the hardships which she endures, and the situation of the baby Cezannes in their hermetically sealed environments, she recounts the facts with interest, but simply doesn't understand why anyone would see those facts as providing any reason—even reasons outweighed by other reasons—to abandon her project.

Let's assume that the Aesthete has freedom of will—on what grounds, after all, could it be denied? The Aesthete is surely self-expressive—a whole range of attitudes and dispositions that are very particular to her are expressed in her choices. But does she express those aspects of herself that are thought to be expressed by the Dutiful? No. The Aesthete exhibits nothing at all to suggest that she even possesses whatever dispositions, attitudes and states are required to recognize and respond to value appropriately. Thus, if the Dutiful is thought to be self-expressive, she is not self-expressive in the same sense as the Aesthete. They have no form of self-expression in common; they don't express the same thing about themselves in their choices nor do they enjoy the same de-

gree of accordance between their choices and other attitudes (the Dutiful, after all, chooses contrary to her desiderative attitudes, the Aesthete in accordance with them).[17] So, either the Dutiful is a counterexample to the claim that self-expression is necessary for freedom of will, or else the Aesthete is.

Freedom of Will II: The Conditions of Self-Transcendence

In the last section I suggested that attempts to cash out freedom of will in terms of self-expression might come about through reflection on what exactly is appealing about the account of freedom of action sketched earlier. Impressed by the fact that an agent who has freedom of action expresses herself in her conduct, we come to think that freedom of will, too, must consist in some form of self-expression. There is, however, another possible moral to glean from the appeal of the account of freedom of action, a moral that leads us towards a self-transcendence view of freedom of will.

As before, we can start with the following question: Even if freedom of action is not all that is involved in full-fledged freedom, why do we take it to capture any sense in which agents can be free? A possible answer to this question begins with a further question: why is it that dependence on the *will* of the agent, rather than on the agent's desires or whims, constitutes freedom of action? Perhaps the answer is that the will, as opposed to desire or whim, is capable of picking out states of affairs as "to be achieved" by virtue of the objective (or at least inter-subjective) value of those states of affairs. In the sense of "desire" and "will" used here, recall, there are certain rationality conditions on willing that do not govern desire. And, we might say, such rationality conditions apply to willing because willing has a point: when functioning correctly, the will aims us towards states of affairs with a force proportionate to their value. The critical phrase in this last formulation is "when functioning correctly": there are, after all, countless examples of choices that fail to aim at what is really, genuinely, valuable. And, this line of thought continues, perhaps this is just what freedom of the will consists in: the right functioning of the will, where the will functions rightly when it leads us to, and is responsive to, the actual value of chosen states of affairs.

Various theorist have reached this conclusion—although not always by quite this route. In contemporary philosophy, views of this sort have been expressed by Robert Nozick, Susan Wolf, Sarah Buss, Paul Benson[18] and in an interestingly different way by John Fischer and Mark Ravizza.[19,20] And this line of thought has a long and venerable tradition: leanings in this direction can be detected in the views of Aquinas, Descartes, Malebranche, Cudworth, Locke and Leibniz, to name a few.[21] To think of freedom of will along these lines is to think of freedom of will as consisting in a kind of self-transcendence. If one's choices are attuned to the evaluative features of one's surroundings, then one's will is guided by something capable of providing a better grounding for choice than can be provided by oneself. This is not to say that the will is not also guided by oneself—one may have an interest, either because of one's desires

or wishes or values, to be disposed to choices appropriate to the value of one's circumstances, or, perhaps, as discussed above, it is simply not possible to be responsive to value in one's choices without also responding to the exercise of certain capacities such as capacities for the recognition of the good. In either event, choices that are responsive to value might also be self-expressive, but what makes them instances of the exercise of freedom of will is, rather, that they are attuned to the facts about value, facts that are not reducible to facts about oneself.

As in the case of the conditions of self-expression, satisfaction of the conditions of self-transcendence—whatever they are—is neither necessary nor sufficient for freedom of will. The following case is a counterexample to the necessity claim: "The Egoist" calmly and coolly assesses the value of the features of his circumstances and the likely results of his actions and chooses to act so as to further his own situation and satisfy as many of his own desires as possible, even if that involves trampling on the needs of others or in some other way realizing states of affairs that are intrinsically disvaluable. It seems clear enough that the Egoist possesses (or could possess) freedom of will. Does he transcend himself in his choices? Those who take satisfaction of the conditions of self-transcendence to be necessary for freedom of will are committed to saying "yes". But the claim that the Egoist is self-transcendent is only plausible when an attenuated conception of self-transcendence is invoked, a conception that seems to collapse into little more than an equation between self-transcendence and the necessary conditions of freedom of will. If the determination to satisfy one's own needs at the expense of others counts as self-transcendence, as responsiveness to the evaluative facts, then self-transcendence is just another word for freedom of will.[22]

The argument against the claim that self-transcendence is sufficient for freedom of will is somewhat more complicated. The trouble is that it is possible for agent's dispositions to recognize and respond to value to be activated by a manipulator in order to serve the manipulator's particular ends. This is so in a wide range of cases of coercion. To take a mundane case, if someone holds a gun to your head and thereby induces you to choose to give her money, she does so by making so choosing the best of your options, and thereby activating your ability to recognize which of your options is best and choose accordingly. But she thereby undermines your freedom of will, without undermining your self-transcendence: you recognize which features of your circumstances rightly provide reason for choosing as you do—most notably the manipulator's firm intention to kill you should you heroically refuse—and you choose in a way appropriate to those reason-giving features. Such cases, then, appear to be cases of agents who are self-transcendent, but lack freedom of will.

Just as the self-expression theorist, when faced with cases of brainwashing, felt the need to resist the first analysis of such cases by claiming that, in fact, they do not exemplify self-expression, the self-transcendence theorist will want to resist the natural conclusion to be drawn from coercion cases by claiming that they do not exemplify self-transcendence. A natural way to defend such

a view is by defining self-transcendence in such a way as to rule out the possibility that the facts about value can be under the control of a manipulator while the agent is still self-transcendent. That is, one might say, to be self-transcendent is not to respond to the evaluative facts as they are, given the actions of a manipulator, but to respond to the evaluative facts as they would be in the absence of the manipulator's manipulations. The trouble with this answer is that it relies on an *ad hoc* distinction between the influence that a manipulator can have on the evaluative facts and the influence that other, quite random, forces can have. I explain.

Imagine a pair of cases: in the first, an agent faces a terrible result if she does not choose to A because a manipulator has promised to bring the terrible result about should she fail to so choose; in the second, the very same terrible result will come about if she fails to choose to A, but the result is assured because of facts quite indifferent to the conduct of the agent. In the first case, for instance, an evil force will cause an agent to be crushed by an avalanche should she choose to take the right fork, while in the second, the avalanche will crush her in just the same way should she make such a choice only because of the precarious position of the snow. According to the line of thought under discussion, if both agents choose to take the left fork for the reasons supplied by the manipulator or the precarious position of the snow, respectively, then the first is not self-transcendent, while the second is. But why? What is it about the influence of another person which takes away the kind of responsiveness to the evaluative facts enjoyed by the second agent?

Perhaps there is an answer to this question, but notice that any answer to it has a very difficult obstacle to overcome: while there might be something freedom-undermining about the influence of another person which could not be duplicated by indifferent forces—perhaps, for instance, the crucial fact is that manipulators, unlike random forces, don't just aim at, but also *track* the compliance of their victims—it isn't clear that any feature of manipulators could be found that isn't present also in cases in which we comply with the wishes of others and thereby reap advantage. To comply with a threat is to be unfree, to comply with an offer is not. The self-transcendence theorist, then, must be able to make both of two distinctions: a distinction between the effect on freedom of will of persons, on the one hand, versus natural forces, on the other; or a distinction between the effect on freedom of will of those who issue threats, on the one hand, and those who issue offers on the other. It isn't clear that a self-transcendence theorist can draw the first distinction without having trouble drawing the second, or draw the second without having trouble drawing the first.

Freedom of Will and the Nature of Evaluative Concepts

So, I take myself to have shown that neither self-expression nor self-transcendence is either necessary or sufficient for freedom of will. But there is something a little peculiar going on. The Dutiful seems to possess freedom of will by virtue of the fact that she is self-transcendent—that would seem to sug-

gest that self-transcendence is sufficient for freedom of will, at least in that case. Similarly, the Egoist seems to possess freedom of will because she is self-expressive, and that would seem to suggest that self-expression is sufficient for freedom of will in her case. Further, in both cases, the relevant feature—self-expression or self-transcendence—seems necessary to the freedom of the individual described: if the Dutiful were to choose against all of her inclinations and desires but, at the same time, to fail to choose what she chooses because of the value of that choice, she would not be free; her choices would be unguided, they would be without satisfactory explanation. Similarly, if the Egoist consistently chose that which furthered her welfare but failed to express anything about herself, she would be, merely, a self-preservationist automata: a creature programmed to pursue her own welfare, but not because of anything deep or important about herself.

What is going on? There are a couple of possibilities: perhaps, we have yet to identify the right feature of agents by virtue of which they are free; or, perhaps, freedom of will must be analyzed as some complicated combination of conjunctions and disjunctions of the conditions of self-expression and the conditions of self-transcendence. Nothing that I've said rules out either of these answers. But there is some reason to think a third answer to be the right one: perhaps freedom of will is a thick evaluative concept.

Consider, first, an uncontroversial example of a thick concept: the concept of pretentiousness in art. We can imagine a series of descriptive features which contribute to, say, a novel's pretentiousness. For instance, if the dialogue in a novel is ponderous—if, that is, characters are frequently making speeches that consist of little more than statements of their philosophical beliefs or their stances on broad moral issues—this might contribute to the novel being pretentious.[23] A novel can be pretentious even if its dialogue is not ponderous and it can have ponderous dialogue and still avoid being pretentious—that is, ponderous dialogue is neither necessary nor sufficient for being a pretentious novel. But, nonetheless, the ponderousness of the dialogue is one of the factors that can be fairly appealed to in an argument that a particular novel is pretentious. Pretentiousness is a "thick concept": it is largely descriptive, but it also imputes to that to which it applies a particular form of disvalue. Determining whether or not the concept of pretentiousness applies to a novel is a matter of evaluatively weighing the novel's descriptive features in order to determine whether or not, taking into consideration the evaluative impact of each of the features, the novel possesses the particular form of disvalue that pretentious novels possess.

Freedom of will, I suggest, is a concept on the model of the concept of pretentiousness. While there are certain descriptive features that agents possess or lack that contribute to their having or lacking freedom of will—in particular, those features that contribute to their being self-expressive or self-transcendent—determining whether or not an agent possesses those features is not enough for determining whether or not the agent has or lacks freedom of will; we must also determine whether or not, by virtue of the possession of the particular combination of relevant descriptive features by the agent, the

series of factors that contribute, causally, to the production of the agent's choice possess or lack value.

I am suggesting, then, that the agent who has freedom of will has choices that come about through worthwhile processes, processes possessing a certain kind of value; she approaches, thereby, an aspect of what we might call "agency at its best". If, in a particular circumstance, the attainment of self-transcendence requires the loss of some degree of self-expression, or vice versa, then it may be impossible for an agent to have all of the descriptive features relevant to freedom of will. But such an agent can still attain freedom of will if, ultimately, by evaluatively weighing the particular degree to which she is self-transcendent at the expense of being self-expressive (or self-expressive at the expense of being self-transcendent) her choice-making processes realize, on balance, the appropriate form of value. So, an agent has freedom of will when there is positive value to her choice-making mechanism, where the value of that mechanism is assessed by taking into consideration the value of both the degree to which she is self-expressive and the degree to which she is self-transcendent.

This account can be formalized by specifying when it is that a particular feature, F, of an agent's circumstances or psychology, which causally contributes to an agent making the choice that she makes, makes a positive or negative contribution to the agent's freedom of will:

> F, a feature of an agent S's circumstances or psychology, makes a positive (negative) contribution to S's freedom of will iff F's presence contributes more positively (negatively) to S's condition evaluated with respect to S's self-expression or self transcendence than F contributes negatively (positively) to S's condition evaluated with respect to the other trait.[24]

Under this view, we can come to a judgment about an agent's freedom of will by applying this test to the conjunction of all the various factors that contribute causally to her choice. So, an agent possesses freedom of will without qualification, when F contributes positively to her freedom of will, and F consists of the conjunction of all the relevant features of her circumstances and psychology.[25]

Why should we think this story might be right? The reason is that the concept of freedom of will functions in various ways that are best explained if freedom of will is a thick evaluative concept. Notice that we learn a lot about a person when we learn what she takes to undermine freedom and what she takes to be irrelevant to freedom. Political conservatives tend to be unwilling to withhold responsibility from a person just by virtue of the fact that that person is a victim of childhood abuse, or an addict, because they tend to think that abuse or addiction do not detract from freedom. Political liberals, on the other hand, tend to be willing to take a wider class of appeals as legitimately undermining the appropriateness of moral censure. We might wonder why it is that one's political and moral stances should have any influence at all on one's judgments of freedom or unfreedom. Whether or not an agent is free is not a political issue—although it has political repercussions in particular cases—and so liber-

als and conservatives shouldn't disagree about who is free and who not; they should disagree only about, say, whether or not an agent needs to be free in order to be justly punished by the state, an argument that might turn on the nature and purpose of punishment by the state, but not on the nature of freedom. If an agent's freedom is like any other descriptive trait of the agent—if the question of whether or not an agent is free has the same status as the question of whether or not the agent was, say, born in Alaska—then disagreements over value should not have an impact on rational assessments of an agent's freedom.

But such disagreements do have an impact on the rational application of the concept of freedom of will. Perhaps this is precisely because whether or not an agent has or lacks freedom of will is, in part, a question of value; answering this question—offering a judgment with respect to an agent's freedom of will—requires offering an assessment of the value or disvalue of the various aspects of her circumstances and psychology which contribute to the production of her choice. There is no reason to expect that such an assessment should, or even could, proceed independently of our evaluative dispositions and attitudes.

It is a commonplace of aesthetic and moral evaluation that the degree of value that we fairly place on that which we are assessing is, in part, a function of what we take to be reasonable to expect from it. It is no fair criticism of a circus that it fails to probe into the nature of the human condition; it is no fair criticism of a small child that she fails to lend support to her parents during their divorce. This fact, I claim, together with the view of freedom I'm suggesting, helps us to explain our intricate intuitions concerning the effect of childhood trauma on the freedom of the adult. A story about the childhood trauma endured by a seemingly wicked person can alter our evaluative tendencies, and thereby alter our judgment about the agent's freedom of will, by influencing our expectations regarding the degree of self-transcendence (or, in some cases, self-expression) that that agent might hope to have.

For instance, imagine that we are given the details of a person's crime and shown that that person knew the evaluative facts and represented them appropriately in her desires, but responded to those facts in a way that seemed to involve the taking of evil for good.[26] Such an agent fails to be self-transcendent in her choices—her choices are inappropriate to the evaluative facts—but there is no reason to think that she is not self-expressive: we can imagine, for instance, that she wholeheartedly and unreservedly does evil, that it is really *her* doing it. Now imagine two different cases. In the first we are given a further story about the loving and supportive environment in which this criminal grew up, and the early penchant that she showed for the infliction of pain that eventually developed into the enraged behavior that has characterized her adulthood. In the second we are given a further story about the terrible abusive circumstances of her upbringing and the prolonged agonies with which she has been inflicted. Given the first agent's early dispositions and tendencies, there is little reason to ever have expected her to get any closer to self-transcendence than she has, and hence the only relevant factors in evaluating the impact of

her circumstances and psychology on her freedom of will are their impact on her degree of self-expression. But they have not contributed negatively to her tendency towards self-expression and so we judge her to have freedom of will. Or, to put the point slightly differently, when we see that she was not *turned* into a monster, but began with tendencies in that direction, it seems that the features that led to her criminal choices did not take anything from her, for we cannot reasonably expect her to have better dispositions towards choice in this respect. Neither have the circumstances of the second agent impugned her ability for self-expression: she is a monster and her monstrous qualities are expressed in her criminal choices. But in the case of the second agent, we think that she could have been self-transcendent; we think that she could have responded to value appropriately if only she had not endured the childhood trauma she endured. It starts to seem, then, that in the case of the second agent the abuse she has endured has taken something from her that she very well could have had: a genuine interest in and attraction towards what is actually good, a form of self-transcendence.[27] The circumstances of the second agent, then, seem to have taken freedom of will from her, for the disvalue of the impact of those circumstances, when assessed with respect to self-transcendence, has outweighed their value with respect to self-expression, because, as we learn from the story of her childhood trauma, there was more for her to lose than the first agent ever could. Various pieces of information—in this case information about the childhood histories of agents—effect our evaluative stances by setting baselines for evaluative judgment, and thereby effect our assessments of freedom of will.

Given the diagnosis of our intuitions about childhood trauma that I am offering, it might be objected, genetic flaws that give rise to, say, violent adult behaviors would not detract from freedom of will. But this seems wrong. Wouldn't we be more likely, the objection goes, to think someone's choice to perform a violent act to be made without freedom of will when that choice is the consequence of flaws in genetic make-up? This objection assumes, however, that the self is not something separate from those aspects of oneself that are dictated by one's genetic make-up. If genetic make-up is something that is imposed on the agent, in something analogous to the way in which childhood trauma is imposed on the agent, then our intuitions with regard to its impact on freedom can be analyzed in just the same way as childhood trauma: we will be likely to see it as detracting from freedom of will since it takes something away from the agent that she might have otherwise had (namely self-transcendence). However, if genetic make-up is constitutive of the self—if there is really no meaning to the thought that one is acted on by one's genetic make-up since there is no self to be acted on prior to the having of some genetic make-up—then it is not possible to think of genetic make-up as something that detracts from freedom of will. In this event, it doesn't detract from self-expression and it can only be thought to detract from self-transcendence if we can meaningfully imagine the very same agent as having a different genetic make-up. This is really a point about essential properties: essential properties of agents cannot be meaningfully examined for their impact on freedom of will, since assess-

ments of freedom of will require evaluative comparison between an agent who has and an agent who lacks the relevant property. When the property is essential, this comparison cannot be meaningfully made.

How do we go about weighing the evaluative importance of self-expression and self-transcendence, in a particular case, in order to decide whether or not and to what degree an agent possesses freedom of will? Unfortunately—and this is why this paper offers only a *strategy for developing* a view of freedom of will rather than a full-blooded view of its own—I cannot really answer this question here. (How do we go about weighing, say, ponderousness and breadth of subject matter in determining the pretentiousness of a novel? The answer would require an essay of its own.) I take myself to have provided an account of which features are relevant to the application of the concept of freedom of will, and some reasons for thinking that their relevance is evaluative. But this leaves open the question of how, precisely, they are to be weighed, and without such an account it is impossible to say, in particular cases, which agents have freedom of will and which lack it.

Evaluation often proceeds in front of a background of purposes: when engaging in the kind of evaluation typical of legal thought, for instance, we keep an eye towards maximizing two potentially conflicting goods: the good of society and the good of individual members of the society. This complicated dual interest influences our judgments of legal responsibility.[28] When making certain sorts of aesthetic judgments—those typical in the criticism of art—we enter into the evaluative process with an interest in certain sorts of pleasure that can be found from the engagement with art objects, and which might, for instance, lead us to ignore the social good or evil consequences of the art object. We can expect evaluative judgments to vary in so far as we enter into the evaluative process with varying aims and interests. And, to the degree that this is so, what has been suggested here is that we can expect our judgments with respect to freedom of will to vary with our evaluative purposes and interests. We can expect, for instance, that our judgment that a particular criminal meets or fails to meet *mens rea* criteria in the criminal law to be different from the judgment that we make with respect to the very same agent's freedom of will when we are thinking of moral, rather than legal, responsibility. Depending on our purposes in inquiring about a particular agent's freedom of will, we may weigh self-expression or self-transcendence to different degrees in coming to our all things considered judgment of the value of the impact of the agent's circumstances and psychology on her will. If the rather tentative suggestion being made here is right, then an agent's freedom of will is to be judged by appropriately situated judges rather than metaphysicians.[29]

Notes

1. Sometimes thoughts of this sort are expressed not by appealing to what we choose, but, instead, by appealing to what we want or what we desire. However, as long as we take the will to be a distinctive capacity, different from the capacity for desire

(as I do—see the later remarks in the main text), there are obvious counterexamples to formulations in terms of desires or wants. I express this familiar thought in terms of choice, then, since I take it to be the strongest formulation.

2. For examples of theorists who have followed each of these various directions of thought, see the later sections entitled "Freedom of Will I: The Conditions of Self-Expression" and "Freedom of Will II: The Conditions of Self-Transcendence".
3. See Wolf, *Freedom Within Reason*, p. 72.
4. One of the earliest usages of the term "thick concept" is in Williams, *Ethics and the Limits of Philosophy*, p. 129.
5. It is in this sense that Michael Smith uses the term desire; cf *The Moral Problem*, pp. 7–8. See also G. F. Schueler, *Desire: Its Role in Practical Reason and the Explanation of Action*, especially the introduction and pp. 29–41, for a useful discussion of this and another conception of desire.
6. We sometimes use the term "desire" to refer not to an occurrent mental state but to a disposition to be in a certain occurrent state in certain circumstances. This is what we mean when we say, of the sleeping child, "She want to please her Daddy." We don't think she wants this right this second, but would in certain appropriate circumstances. As I am using the terms, the dispositional state being referred to here is not a desire but merely a disposition to have certain desires.
7. In his excellent book *Intention, Plans, and Practical Reason*, Michael Bratman identifies a series of norms of rationality governing intentions and other acts of will. See, especially, chapter 2.
8. Hobbes' view of freedom is expressed most clearly and explicitly in his essay "Of Liberty and Necessity". Following the publication of this essay, Hobbes engaged in an extensive correspondence over the issues with Bishop Bramhall. Hobbes's essay and much of the correspondence between Hobbes and Bramhall appears in Vere Chappell's *Hobbes and Bramhall on Liberty and Necessity*.
9. This condition could be formulated, with a little ingenuity, in any tense. I use the present tense here only for convenience.
10. Alternatively, the second of these conditional might be formulated as follows: (2') If she chooses to refrain from A, she will refrain from A. What makes (2) preferable to (2') is that we can construct cases in which an agent can only avoid A-ing by making no choice with respect to A. If, for instance, whenever I choose to refrain from tripping, I get so nervous that I trip, then I cannot avoid tripping by choosing to refrain from tripping. But I might, nonetheless be free, in some weak sense, with respect to tripping since I can avoid tripping by thinking of other things and thereby making no choice in favor or against tripping.
11. The objection was posed to Hobbes by Bramhall. See *The English Works of Thomas Hobbes*, v. 5, cf. pp. 40–41. Or Chappell, *Hobbes and Bramhall on Liberty and Necessity*, pp. 43–44. Bramhall puts the objection slightly differently than I do in the main text; but, Bramhall does emphasize that, according to Hobbes' definition of freedom, even creatures such as madmen and children—creatures who have no control over what they choose—are still free.
12. This objection is no different from the objection posed to conditional analyses of "can" by Roderick Chisholm. Chisholm claims that for it to be the case that an agent can do something it must be the case that she can choose to do it, and points out that advocates of the conditional analysis don't require this further condition (see Chisholm, "Human Freedom and the Self", p. 27). In my opinion, Chisholm is simply sensing that it is possible to undermine freedom by tampering with the way

an agent chooses without thereby tampering with the way the agent acts given her choices.

13. It is possible that views of this sort are not so far away from agent-causal theories of the sort that begin with Thomas Reid and continue in the work of Roderick Chisholm and, more recently, Timothy O'Connor and Randolph Clarke. (Reid, *Essays on the Active Powers of Man*, especially Essay IV, chapters 1 and 2; Chisholm, "Human Freedom and the Self"; O'Connor, "Agent Causation" and Clarke, "Toward a Credible Agent-Causal Account of Free Will".) The agent causal theorists do not describe their project as one of unpacking the conditions of self-expression, but, nonetheless, their project can be construed in this way. The agent causalists are concerned that if events or states, rather than agents, are the causes of conduct, then the critical fact about agents by virtue of which they are agents will remain unexpressed in conduct. That is, they take capability for action independent of the causal influence of any particular states or events to be the crucial agency-defining feature and think that choices that are not expressive of this feature—by virtue of being caused by the agent herself—are thus not free. Thus, the Frankfurtian and Reidian projects all aim to capture some sense in which our choices can succeed or fail to be self-expressive. The views differ in that Frankfurt (and those who offer related theories) attempt to capture the relevant sort of self-expression by appealing to certain crucial elements of the agent's psyche—in Frankfurt's original paper he appealed to what he called "second-order volitions"—and the relationship that they bear to other elements such as choices. The agent causal theorists, on the other hand, try to capture the special kind of self-expression by appealing to a special kind of cause: an agent, irreducible to any particular features of the agent.
14. The distinction between Frankfurt's willing addict and a recreational drug-user—someone who is not addicted but acts on a desire to take a drug and has an appropriate higher order attitude in support of that desire—must be made by appeal to dependence between the higher order attitude and the effective first order desire. The recreational drug-user's desire for the drug would not be effective were she not to have the higher order attitude in favor of it; not so for the willing addict. This difference is a difference in self-expression, and thus the degree of self-expression enjoyed by an agent, under Frankfurt's view, is a function, in part, of the counterfactuals which the actual causal sequence leading to choice possesses.
15. See, for instance, Frankfurt's "Identification and Externality", "Identification and Wholeheartedness" and "The Faintest Passion".
16. See, for instance, Friedman, "Autonomy and the Split-Level View", Neely, "Freedom and Desire", Stump, "Sanctification, Hardening of the Heart and Frankfurt's Concept of Free Will", Watson, "Free Agency" and Young, "Autonomy and the 'Inner Self'".
17. It remains open to the self-expression theorist to claim that, despite appearances to the contrary, the Dutiful and the Aesthete have some form of self-expression in common, or to claim that the requisite form of self-expression is disjunctive (an agent must express herself either in the way the Dutiful does or else in the way the Aesthete does, or...), but this further squirming would seem to have diminishing returns in the form of loss of unity in the concept of self-expression.
18. Nozick, *Philosophical Explanations*, pp. 317–362; Wolf, *Freedom Within Reason*, especially chapter 4; Buss, "Autonomy Reconsidered", and Benson, "Freedom and Value".

19. See Fischer, *The Metaphysics of Free Will*, especially chapter 8, and Fischer and Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*. Fischer and Ravizza see the right functioning of the will, to use my terminology, as consisting in choice-producing mechanisms that are "reasons-responsive" in a special sense that they make an effort to define. Their central idea, however, involves thinking of freedom of will as consisting in routes to choice that are grounded in and responsive to features of circumstances that are relevant to making right choices, where "right choices" are just those supported appropriately by reasons. Put in this way, their view does seem to involve something like the strategy that I am discussing.
20. The view expressed in the works of Michael Smith (both by himself and with Jeannette Kennett and Philip Pettit—see Pettit and Smith, "Backgrounding Desire" and "Freedom in Belief and Desire", and Kennett and Smith, "Frog and Toad Lose Control", and Smith, "A Theory of Freedom and Responsibility") probably also belongs in this category. However, Smith et al's insistence that the aspect of their view which sounds like an account of self-transcendence is not rightly called autonomy, but, instead "orthonomy", suggests that they are not thinking of the view in quite the way that I am suggesting.
21. This is obviously not the place for a full discussion of the views of these various figures. A small sampling of texts that might begin to justify my sweeping historical claim here: Aquinas, especially *Summa Theologica* I.q 83, I-II.q 8, I-II.q 10; Descartes, especially the Fourth Meditation, AT VII 57–58; Malebranche, *Treatise on Nature and Grace*, especially Discourse III, sections 8–10; Cudworth, *A Treatise Concerning True and Immutable Morality*, notably pp. 26–27, and *Treatise of Free Will*, especially chapter 8; Locke, *An Essay Concerning Human Understanding*, II.XXI especially II.XXI.48–50; Leibniz, *Theodicy*, especially part I section 45, part II sections 228–235, and part III sections 310 and 319. For a discussion of the role that this line of thought plays in Locke's view of free agency, see my *Liberty Worth the Name: Locke on Free Agency*, especially chapter 1.
22. It might be argued, I suppose, that self-transcendence doesn't require actually choosing rightly, but, rather, choosing in a way that correctly takes into account where the good lies. Under this conception of self-transcendence, even the Egoist can be self-transcendent since he might carefully take into account what it is actually best for him to do and simply decide to do what furthers his own interests. He is responsive to the evaluative facts on this model, he just doesn't respond to them in the ideal way.

    This conception of self-transcendence will only seem a viable theoretical option to those who take a strictly non-internalist conception of value properties. That is, those who think that to judge a particular possible one of one's actions to be morally valuable is to be motivated to perform it don't allow for the possibility that one could correctly judge a particular action to be the best of one's options and not be motivated to perform it more strongly than one is motivated to perform any other action. Thus, moral judgment internalists cannot accept the account of self-transcendence under discussion here. However, there may be room for those who reject moral judgment internalism to dispute the claim that the Egoist is a counterexample to the necessity of self-transcendence for freedom.
23. We might worry that the concept of ponderousness is itself a thick concept. Let's assume that it is not. Let's assume, that is, that something like the descriptive definition of ponderousness given in the main text is correct.

24. Notice that it is unclear in the absence of further argument whether or not determinism undermines freedom on this account. In principle, at least, it could be argued that deterministic choice-producing mechanisms possess disvalue, that their effect on self-expression and self-transcendence are, in the end, evaluatively negative. I myself do not know of a satisfying argument to this effect, but I leave it to others to debate the issue. Notice, however, that if it can be shown that the truth of determinism does not undermine the possibility of choice-producing mechanisms that allow either self-expression or self-transcendence, or both, then the theory will turn out to be compatibilist.
25. In "Self Deception and Responsibility for the Self", Stephen White reaches the following conclusion regarding the impact of self-deception on responsibility:

> We must drop the assumption that our practices of ascribing responsibility could have a justification in which discriminations in our ascriptions to different subjects are justified by differences in the intrinsic properties of those subjects' psychologies. (p. 478)

While the suggestion I am making is not as strong as this—for all that has been said, the value of the mechanism through which a person's choice comes about might supervene entirely on "intrinsic properties" of that mechanism—there is some affinity between White's view and the view I am suggesting.
26. An example, perhaps, of such a case is discussed at length by Gary Watson in his enormously thought-provoking article "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme".
27. Some readers may be disturbed by my usages of "could" here. After all, they might say, exactly what we are trying to understand when we give accounts of freedom is in what sense, exactly, alternative scenarios need to be available to agents for freedom. But, the sense of "could" I am invoking is very wan indeed and gets nowhere near amounting to the "could" analyzed by an adequate account of freedom. In particular, there is no reason to think that the agent herself could have done anything to bring it about that she be closer to being self-transcendent; it is simply the case that under different early influences, she would have come to have a different character, and not through any action on her part.
28. See Feinberg, "Problematic Responsibility in Law and Morals" for a marvelous discussion of the ways in which the purposes which we have when assessing legal responsibility differ from those involved in assessing moral responsibility.
29. Thanks to Sarah Buss, Vere Chappell, Phillip Clark, Andrew Eschelman, John Fischer, Paul Hoffman, Elijah Millgram, John Perry, Vance Ricks, Jennifer Rosner, Marleen Rozemond, Kadri Vihvelin, and Gary Watson for comments on this paper or its recent ancestors. Portions of this material were presented to the philosophy departments at the University of California at Irvine and Arizona State University. In both cases I received valuable comments. Special thanks to Michael Bratman for his infinite willingness to re-read, and for always seeing exactly what needs to be fixed.

References

Aquinas, T. [1270] (1945) *Basic Writings of Saint Thomas Aquinas*, Random House, New York.
Benson, P. (1987) "Freedom and Value", in *Journal of Philosophy*, pp. 465–486.

Bratman, M. (1987) *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge.
Buss, S. (1994) "Autonomy Reconsidered", in *Midwest Studies in Philosophy*, v. 19, pp. 95–121.
Chappell, V. (ed.) (1999) *Hobbes and Bramhall on Liberty and Necessity*, Cambridge University Press, Cambridge.
Chisholm, R. (1964) "Human Freedom and the Self", in Watson, G. (ed) *Free Will*, pp. 24–35, Oxford University Press, Oxford, 1982.
Clarke, R. (1993) "Toward a Credible Agent-Causal Account of Free Will", in O'Connor, T. (ed) *Agents, Causes and Events: Essays on Indeterminism and Free Will*, pp. 201–215, Oxford University Press, New York, 1995.
Cudworth, R. [1731] (1996) *A Treatise Concerning True and Immutable Morality*, Hutton, S. (ed), Cambridge University Press, Cambridge.
Cudworth, R. [1838] (1996) *A Treatise of Freewill*, Hutton, S. (ed), Cambridge University Press, Cambridge.
Descartes, R. (1984) *The Philosophical Writings of Descartes*, v. 1–2, Cottingham, J., Stoothoff, R. and Murdoch, D. (trans.), Cambridge University Press, Cambridge.
Feinberg, J. (1962) "Problematic Responsibility in Law and Morals", in *Doing and Deserving: Essays in the Theory of Responsibility*, pp. 25–37. Princeton University Press, Princeton.
Fischer, J. (1994) *The Metaphysics of Free Will*, Blackwell, Oxford.
Fischer, J. and Ravizza, M. (1998) *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, Cambridge.
Frankfurt, H. (1971) "Freedom of the Will and the Concept of a Person", in *The Importance of What We Care About*, pp. 11–25. Cambridge University Press, Cambridge.
Frankfurt, H. (1976) "Identification and Externality", in Rorty, A. (ed),*The Identities of Persons*, University of California Press , Berkeley, 1976.
Frankfurt, H. (1987) "Identification and Wholeheartedness", in Fischer, J. M. and Ravizza, M. (eds) *Perspectives on Moral Responsibility*, pp. 170–187. Cornell University Press, Ithaca.
Frankfurt, H. (1992) "The Faintest Passion" in *Proceedings and Addresses of the American Philosophical Association*, v. 66.
Friedman, M. (1986) "Autonomy and the Split-Level View" in *Southern Journal of Philosophy*, v. 24, pp. 19–35.
Hobbes, T. (1646) "Of Liberty and Necessity", in *The English Works of Thomas Hobbes*, v. 4, Scientia Verlag Aalen, Germany.
Hobbes, T. (1648) "The Questions Concerning Liberty, Necessity and Chance, Clearly Stated and Debated Between Dr. Bramhall, Bishop of Derry, and Thomas Hobbes of Malmesbury", in *The English Works of Thomas Hobbes*, v. 5, Scientia Verlag Aalen, Germany.
Kennett, J. and Smith, M. (1996) "Frog and Toad Lose Control", in *Analysis*, v. 56, n. 2, pp. 63–73.
Leibniz, G. W. (1996)*Theodicy*, Huggard, E. M. (trans.), Open Court, La Salle.
Locke, J. [1690] (1975) *An Essay Concerning Human Understanding*, Clarendon Press, Oxford.
Malebranche, N. [1680] (1992) *Treatise on Nature and Grace*, Riley, P. (trans. and ed.), Clarendon Press, Oxford.
Mele, A. (1992) "Recent Work on Intentional Action", in *American Philosophical Quarterly*, v. 29, n. 3, pp. 199–217.
Moore, G. E. (1903) *Principia Ethica*, Cambridge University Press, Cambridge.
Neely, W. (1974) "Freedom and Desire", in *Philosophical Review*, v. 83, pp. 32–54.
Nozick, R. (1981) *Philosophical Explanations*, Harvard University Press, Cambridge.
O'Connor, T. (1995) "Agent Causation", in O'Connor, T. (ed) *Agents, Causes and Events: Essays on Indeterminism and Free Will*, pp. 173–200, Oxford University Press, New York, 1995.
Pettit, P. and Smith, M. (1990) "Backgrounding Desire", in *Philosophical Review*, v. 99, pp. 565–592.
Pettit, P. and Smith, M. (1996) "Freedom in Belief and Desire", in *Journal of Philosophy*, v. 93, pp. 429–449.
Reid, T. (1788) *Essays on the Active Powers of Man*, Lincoln-Rembrandt Publishing, Charlottesville.
Schueler, G. F. (1995) *Desire: Its Role in Practical Reason and the Explanation of Action*, MIT Press, Cambridge.

Smith, M. (1994) *The Moral Problem*, Blackwell, Oxford.
Smith, M. (1997) "A Theory of Freedom and Responsibility", in Cullity, G. and Gaut, B. (eds) *Ethics and Practical Reason*, Oxford University Press, Oxford, 1997.
Stump, E. (1993) "Sanctification, Hardening of the Heart and Frankfurt's Concept of Free Will" in Fischer, J. M. and Ravizza, M. (eds) *Perspectives on Moral Responsibility*, pp. 211–234. Cornell University Press, Ithaca, 1993.
Watson, G. (1975) "Free Agency", in Watson, G. (ed) *Free Will*, pp. 96–110, Oxford University Press, Oxford, 1982.
Watson, G. (1977) "Skepticism about Weakness of Will", in *Philosophical Review*, pp. 316–339.
Watson, G. (1987) "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme", in Fischer, J. M. and Ravizza, M. (eds) *Perspectives on Moral Responsibility*, pp. 119–150. Cornell University Press, Ithaca.
White, S. (1988) "Self Deception and Responsibility for the Self", in McLaughlin, B. P. and Rorty, A. O. (eds) *Perspectives on Self Deception*, pp. 450–484. University of California Press, Berkeley.
Williams, B. (1985) *Ethics and the Limits of Philosophy*, Harvard University Press, Cambridge.
Wolf, S. (1990) *Freedom Within Reason*, Oxford University Press, Oxford.
Yaffe, G. *Liberty Worth the Name: Locke on Free Agency*, forthcoming, Princeton University Press, Princeton.
Young, R. (1980) "Autonomy and the 'Inner Self'" in *American Philosophical Quarterly*, v. 27, pp. 35–43.