5-2010

Clarifying Intuitions about Moral Responsibility and the Self

Daniel Carl Homer
College of William and Mary

Clarifying Intuitions about Moral Responsibility and the Self

A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of the Arts in Philosophy from
The College of William and Mary

by

Daniel Carl Homer

Accepted for _____
(Honors, High Honors, Highest Honors)

_____
Matthew Haug, Director

_____
Neal Tognazzini

_____
Michael Green

Williamsburg, VA
April 22, 2010

Introduction

Despite the significant differences between the positions of the libertarian, compatibilist and pessimist[1], they all seem to start with the same set of three intuitions about moral responsibility. Each position places the most emphasis on a different intuition, but each attempts to capture all three in some form. The three intuitions are as follows. If an agent *S* is morally responsible for an action *E*[2], performed at time *t*, then

(1) *S* had control over *E*'s happening at *t*.

(2) *S* was able to do something other than *E* at *t*.

(3) S was morally responsible for the relevant characteristics of her self at *t*.

Those intuitions state necessary conditions for moral responsibility. Of course, even taken together they don't supply a sufficient condition. At the very least, *E* must fulfill some kind of epistemic condition: she must know that what she's doing has moral weight, what other options she had, and so on. Like much of the literature, however, I'm ignoring that condition. My interest is in those three intuitions and their interaction.

Each of the intuitions is pretty vague, but that's not surprising, since they occur to us pre-reflectively and (and least close to) universally. Much of the work in the free will debate has gone to stating the intuitions more precisely and evaluating those new

[1] I use 'pessimist' as Galen Strawson does in "Luck Swallows Everything," *Times Literary Supplement* (June 26, 1998), to mean a person who thinks moral responsibility, at least in any meaningful sense, is impossible.

[2] Some might think that using the term 'action' already assumes certain things about the relationship between the agent and the event, but I don't mean the term in that way. For the purposes of this paper, simply take 'agent' to mean 'the thing with the property of being morally responsible' and 'action' to mean 'the event for which the agent is morally responsible.'

statements. Usually that process brings out differences in people's intuitions and leads directly to the traditional points of disagreement in the debate. For instance, libertarians understand (2) to mean that at *t*, holding fixed the past and the physical laws, not-*E* could have happened. Pessimists usually hold the same. Compatibilists vary on their exact interpretation of (2), but all of them understand it to mean something weaker than the libertarians do (or recognize it as an intuition but reject it as misguided).

Unfortunately, since philosophers move directly from vague points of agreement to specific points of disagreement, it's very hard for anyone to convince somebody who doesn't already share his view. People develop more (and more complicated) arguments and justifications for their position, and others find holes in those arguments, but people rarely seem to change their minds. And why should they? The arguments offered don't actually attack the justification for the position but various accounts of responsibility that support it. Someone might well abandon a particular approach for fleshing out a position and switch to another, but most likely won't switch to a different position altogether. Suppose, for example, that I'm a libertarian and try to articulate my beliefs using Democritus' atomic swerve theory. It doesn't take very complicated arguments to show that that theory is a dead end, though. Among other things, it doesn't seem to leave room for (1). For that reason, I might move to something more sophisticated, like Kane's parallel processing theory. The arguments against Democritus, however, aren't likely to cause me to give up the libertarian position and become a compatibilist.

In order to find arguments for and against adopting libertarianism and compatibilism in the first place, we need to examine the move from the three basic intuitions to those positions. We usually assume that the move is the result of underlying,

irreducible differences in our starting intuitions, so that no position is superior in reference to our intuitions (meaning everyone's intuitions taken together, not an individual's). I think that's a bad assumption. We might be able to state the intuitions more clearly without begging the question against any of the established positions. In other words, we will state the basic components of the intuition (for example, what we mean by 'control' in (1)), as well as the motivation for accepting the intuition (For example, it's not immediately obvious that moral responsibility requires (2), so we must have some reason for thinking it a necessary condition).

If we can do that, we can use the new formulations for two tasks. First, we can see if any of the intuitions are somehow self-contradictory, i.e. if the individual components of the intuition contradict each other. If they are, we can either discard them completely or modify them in a way that dissolves the contradiction. We would need to modify them in a way that still serves the motivation for accepting the intuition, of course. Probably we should favor modification over rejection, simply because all of the intuitions seem fairly hard to give up.

Second, we can see if the intuitions taken together lead to any contradiction. If they do, our course is a little more difficult, because we will probably have to favor one intuition over another. Each camp generally places more importance on different intuitions: compatibilists tend to think (1) most important, libertarians (2) and (3), and pessimists (3)[3]. If we favor one intuition over the others, then, we may start to beg the question in the debate between the camps. If we're lucky, though, we will be able to resolve the contradiction without favoring the motivation behind any of the intuitions,

[3] Most libertarians try to create an account that fulfills (3), while most pessimists spend their time showing that no account adequately does so.

which, remember, we've made explicit. I believe that the intuitions taken together *do* lead to contradiction, at least when all of them are read strongly. I think the strong reading is the most natural for each intuition, so we should see any weaker reading as a revision. I tackle these two tasks in the first chapter.

The above intuitions lie at the heart of the free will debate, but they aren't the only important ones. Participants in the debate often refer to the self and, of course, to action, but no consensus about an acceptable account of the self exists. This can cause problems, since different positions may implicitly rely on different accounts. Further, I haven't seen an account that allows an agent to act and be morally responsible for that action. I offer an account that does in the second chapter.

Before I continue, though, I need to justify a piece of my methodology. I totally ignore the pessimist in this paper, and here's why. We have one more intuition regarding moral responsibility, but it's so obvious that people usually don't even notice it. That intuition is:

(4) Moral responsibility is (logically) possible.

Including (4) begs the question against the pessimist, but I think it's nevertheless justified. (4) seems much more basic than the other three. When we think of moral responsibility, we don't have the same difficulty as we do when we try to think of square circles, or red screeches, or even the largest integer. In that sense, it certainly doesn't seem impossible. In addition, as Peter Strawson points out[4], we possess certain reactive

[4] Peter Strawson, "Freedom and Resentment," in *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001), 187.

attitudes that we often attribute to others (and ourselves) based on their behavior. Most philosophers understand moral responsibility in terms of those attitudes. We clearly assume, not only that moral responsibility is possible, but also that many people possess it. Now, I obviously don't want to rely on popular opinion for justification, but the fact that an assumption of moral responsibility underlies so much of our behavior means that it will be much simpler to revise our understanding of the property than give it up entirely.

More importantly, however, we need to assume (4) in order to meaningfully evaluate the other three intuitions. If we allowed that moral responsibility might be impossible, we would have no reason to assume that its necessary conditions were consistent. In fact, if it were impossible, that would entail that its necessary conditions were inconsistent. Since my entire goal here is to analyze (1) through (3) and attempt to make them consistent, it seems best to simply ignore the pessimist's worry. My goal is to make progress in the debate between the compatibilist and libertarian, who at least agree that moral responsibility is possible. The believers will have to find a different kind of approach to refute the skeptic.

Chapter I: Parsing the Intuitions

In this chapter I consider each of the three intuitions in turn, make explicit the necessary conditions for moral responsibility they entail, and determine if those conditions are consistent. I argue that (1) and (2) are internally consistent, that (3) is internally inconsistent, and that (1) and (2) interact in a way that requires us to understand (2) as a weaker requirement than libertarians typically do. I'll begin, naturally, with (1).

I. The First Intuition

To refresh your memory, the first intuition goes as follows: If *S* is morally responsible for *E*'s happening at *t*, then

(1) *S* had control over *E*'s happening at *t*.

So (1) obviously concerns *control*. What, then, is control? Like just about everything in philosophy, it's a little controversial and a little complicated. Fortunately, I don't have to completely derive a new definition for 'control.' I don't even have to give a complete definition of it. At this point, all I'm concerned with is the portion of a definition that the various parties in the free will debate seem to a) agree upon, and b) implicitly rely upon in their arguments. That kind of control seems like shorthand for three conditions. They are as follows.

(1a) *S caused E* to happen.

(1b) *S intended* for *E* to happen.

(1c) *S* intended to cause *E* to happen in the way, and at the time, she *in fact* caused *E* to happen. [5]

(1a) seems like a fairly obvious condition for responsibility. It's hard to see how an agent could be responsible for an event that she didn't cause. Of course, *S* doesn't have to *directly* cause *E* in order to be responsible for it. For example, assume that Bert wants Ernie to die, but he doesn't want to get in trouble for it. He knows about Ernie's obsession with rubber ducks, so late one night he glues one to the middle of the road in front of Ernie's house. At seven o' clock the next morning, Ernie walks out to get the newspaper. He sees both the duck and a truck bearing down on it, so he rushes into the road to save the fake little waterfowl. The truck honks, but Ernie pays it no mind. He grabs the duck, but the duck is stuck! He pulls frantically on the small rubber bird as the truck's brakes squeal. A few seconds later the truck rolls on, revealing Ernie's grisly end.

Surely we would blame Bert for Ernie's death, despite his indirect method of murder. On the other hand, if Bert had simply *wished* for Ernie to die, or if he had carried out his plan but Ernie had drowned in his oatmeal instead, we wouldn't blame Bert for his death. In those cases, Bert wouldn't have caused Ernie's death, directly or indirectly.

Note that I haven't said anything about the exact causal relation between agent and event. So far we have seen no reason to favor deterministic causation over

[5] These conditions, taken together, come fairly close to the 'standard story of action', as described in J. Michael Velleman, "What Happens When Someone Acts?" *Mind* 101, 403 (1992), 461. However, the standard story has some serious problems and I see no reason to assume everyone in the free will debate implicitly assumes it's a good account. Instead, I'll leave vague exactly how the interactions work and leave a discussion of the standard story for Chapter II.

indeterministic causation, so the compatibilists and incompatibilists are on equal footing[6]. Considering (at least one form of) indeterministic causation provides a good illustration for why (1a) isn't sufficient for control, though.

Suppose Ernie is deciding whether or not to take a bath. Suppose also that Ernie lives in a world where his intentions lead indeterministically to actions, that he decides to take a bath, and that that decision leads in a normal way to an intention to take a bath. Suppose finally that his intention to take a bath leads indeterministically to his singing a song instead. Well, now we have a problem. Ernie caused his singing, but he certainly doesn't seem responsible for it: he was trying to take a bath! If he had instead intended to sing, we would probably hold him responsible for singing. Likewise, if his intention to bathe had actually led to bathing, we would hold him responsible for that. Including (1b), then, seems to solve the problem.

(1a) and (1b) together aren't sufficient to capture the intuition of control, though. Consider the following case: Bert wants to kill Ernie, but he's concerned that people will suspect him of setting Ernie up. So, the day before he intends to kill Ernie, he takes Ernie shopping, hoping people will see the two of them together and on good terms. On the way back from the mall, Bert accidentally runs his car off the road and hits a tree, injuring himself and killing Ernie. Bert killed Ernie, and he had an intention to kill him, so meets (1a) and (1b), but he clearly didn't have control over the way he in fact killed him. Bert certainly doesn't seem morally responsible for Ernie's death; it was an accident, happy though it may be for Bert. Bert's intention needs to match up properly

[6] The *degree* to which we hold people responsible might depend on what kind of causal mechanism is in play, though.

with actual events in order for him to count as morally responsible. (1c) captures that quite nicely.

(1) is consistent with our common practices, but has no further theoretical justification. If someone were to ask why moral responsibility requires the agent to have control over her action, I could only respond, "It just does. That's just part of what moral responsibility is.[7]" Yet it seems well-justified. The burden of proof surely lies with anyone who claims you don't need control to be morally responsible. (2) and (3), as we will see, do require some further justification. It therefore seems that (1) is more basic than the other two, which may provide some reason for preferring it. As I mentioned in the introduction, however, I want to avoid any move that might beg the question in the argument between compatibilists and libertarians, so I won't treat it as prima facie more important.

Now that I've made explicit (1)'s components and (lack of) external justification, I can determine if adhering to it makes moral responsibility impossible. And it pretty clearly doesn't: (1a), (1b) and (1c) don't contradict each other. Here's a demonstration. Form an intention to read the next sentence. When you're done reading this sentence, you'll have fulfilled (1a), (1b) and (1c). Since anything actual is possible, the conjunction of (1a), (1b) and (1c) is possible. (1) represents that conjunction. Considered on its own, then, (1) seems like a reasonable necessary condition for moral responsibility.

[7] I should mention that a growing minority of philosophers disagrees that moral responsibility for an act requires the agent to cause that act. I've chosen to ignore them in this paper, however, because they lie outside the traditional free will debate and, frankly, I'm not sure what to do with them. They reject an intuition that I (and most philosophers) take as basic, so I'm afraid that any debate with them will stall pretty early.

II. The Second Intuition

The second intuition is more contentious than the first, even just with respect to its meaning. Before I talk about that, though, I need to deal with a problem: (2) doesn't properly capture our intuitions about moral responsibility. Again, it says that, if *S* is morally responsible for *E*'s happening at *t*, then

(2) *S* was able to do something other than *E* at *t*.

Unfortunately, (2) falls prey to an obvious counterexample. Imagine that you wake up one morning and don't feel like being at work on time, let's say 8:00 AM. The only way you can make it to work on time is if you leave at 7:30 and drive straight to work. You sleep until 7:45 and show up to work late. Are you responsible for that? It sure seems like it, but (2) says you aren't. At 8:00, when you were still staggering groggily around your kitchen, it was impossible for you to be at work on time. That means that at 8:00, you couldn't have done something other than be late. So, by (2), you're not morally responsible for being late.

Luckily, there's a fairly simple way to get around the problem: make (2) disjunctive, like in the following: If *S* is morally responsible for *E*'s happening, then

(2') At *t, S* could have done something other than *E*, or is morally responsible for the fact that it is impossible for her to do something other than *E*.

Apply (2') to the case above. Assume that you're morally responsible for sleeping in (and it seems you are, given that you knew it would lead to your being late). Your

sleeping in made it impossible for you to do other than be late at 8:00. So you satisfy the second clause of the disjunction, and so meet at least that necessary condition for moral responsibility.

Some might object to adding such a clause on the grounds that it's simply adding an epicycle to the theory, just adding an extra piece to deal with a counterexample. While I typically prefer simplicity and avoid convenient conjunctions and disjunctions, I'm not bothered by this one. Here's why: (2') actually fits with (1) better than (2) does. Remember, (1) requires that *S* cause *E* at *t* and have an intention that *E* happen at *t*. *S* can cause *E* in two ways: directly and indirectly. The causal chain linking *S*'s action to *E* can be any determinate length (including zero, in the case that *S* acts by *E*-ing). Control is preserved along the causal chain.

(2') creates a similar condition. *S* might face a situation such that she cannot do something other than *E* at *t*, but as long as *she* constructed the situation, and is morally responsible for that construction, that doesn't matter. In this case, one might think of *S*'s constructing the restrictive situation as equivalent to *S*'s choosing to *E* at *t*. At the time of her constructing it, she had the ability to do something other than construct it, as stipulated (she is morally responsible for constructing it). With (2'), in a sense, the ability to do otherwise is preserved along the causal chain.

Incompatibilists think we need to read (2') in a way that agents have 'strong' or 'real' alternatives, such that the same person in the same circumstances can do two different things. I can only see two possible reasons for that position. First, maybe they think that moral responsibility just requires the agent to be free from the constraints of the past and the laws of nature in making their decisions. In other words, maybe they think

being free in that way is a basic requirement for moral responsibility, like I (and most people) think being in control is a basic requirement. But having 'strong' alternatives doesn't actually allow that kind of freedom. The past produces the situations in which the agent finds herself and the options open to her in those situations. Indeterminist event-causal and agent causal accounts don't free the agent from the laws of nature; they just replace determinist causation with a different kind.

Second, maybe incompatibilists want 'strong' or 'real' alternatives because of the following fact about moral responsibility: the degree to which we hold people responsible depends on the degree of moral difference between the acts that a situation leaves possible. In other words, the same action has different moral weight in different circumstances. That reason doesn't have the problems the first one has and is, I think, consistent with our intuitions. It may even be a basic intuition.

Imagine Bob is faced with a choice: wink his left eye, kick a puppy, or allow a puppy to die[8]. If he decided to wink, we would probably think him praiseworthy; that choice has the best moral outcome. But imagine he is faced with a different set of choices, wink his left eye, feed some starving children, or stop terrorism once and for all. If he decided wink in those circumstances, we would probably blame him quite a bit. Now, imagine that Bob faces the following choice: wink his left eye, wink his right eye, or blink. In that case, Bob's choosing to wink his left eye bears no moral weight at all, so we probably would neither praise nor blame him. Finally, imagine that Bob chooses from a set of exactly one option: blink his left eye. He doesn't seem praise- or blameworthy for

[8] Assume that those three options cover all the possible outcomes of Bob's choice. If Bob abstains, or sucks his toes, or sings the Canadian National Anthem in falsetto, for instance, assume that his interrogators will kill a puppy, so those actions are equivalent to choosing the third option.

winking his left eye in that circumstance, precisely because he had no choice but to wink his left eye. We have no alternatives with which to compare moral differences.

Of course, options that people can't actually choose don't make any difference in our moral assessment. Suppose that you see a small boy cradling a dead sparrow in his arms. He explains, crying, that he accidentally hit the bird with his baseball and killed it. What's the best thing for you to do, morally speaking? You probably think it's something like 'comfort the boy,' but you're way off. Clearly, resurrecting the sparrow is morally preferable to that! In fact, you'd be even more praiseworthy if you took the opportunity to bring peace to the Middle East, cure cancer and sucker-punch Satan. Since you can't do those things, though, we don't consider them when we judge your action. Ought implies can.

That's the incompatibilists' point: if you can't really do something (they say), it doesn't figure into our moral judgments of you, and in a deterministic world, you can't do anything except what you actually do. Your perception of options is illusory. Given the past and the laws of the universe, you can only really do one thing. And if you can only do one thing, you aren't morally responsible for doing it[9]. (2'), then, is meant to capture the following intuition: you are responsible for *E* happening *instead of* not-*E*, a different thing that could have happened instead. *S* must make the difference between *E*'s occurring and (some alternative to *E)*'s occurring. So where we have '*E*' in (1) and (2'), we should read '(*E* instead of *A*),' where *A* stands for some alternative to *E* available to *S*.

I think they're confused, though.

[9] Compatibilists disagree that you need strong alternatives and argue for a different understanding of (2). I discuss those alternative understandings in Section V, below.

Let's see what (1) looks like, if we make the substitution that (2') seems to warrant, leaving explicit the conditions (1) entails. *S* is morally responsible for *E* (instead of the alternatives)'s happening at *t* only if

(1a') *S* caused (*E* instead of *A*) to happen.
(1b') *S* intended to cause (*E* instead of *A*) to happen.
(1c') *S* intended to cause (*E* instead of *A*) to happen in the way, and at the time, she *in fact* caused (*E* instead of *A*) to happen.

Let's call the conjunction of those three conditions (1'). The problem as I see it is that allowing the agent to have 'strong' or 'real' alternatives makes (1') impossible to fulfill. The libertarian requirement consequently makes moral responsibility impossible. Well, maybe that's too strong. I can at least say with certainty that every event-causal libertarian position I have encountered fails to meet all the necessary requirements for moral responsibility, and fails to do so precisely because of the 'and not-*A*' requirement.

To better show the problem, I'll use a simple event-causal libertarian account as something of a straw man. On this account, the agent considers her various choices, the reasons she has for each one, and, based on those reasons, indeterministically acts in accordance with one of those choices. When I say that she indeterministically acts, I mean that she causes an event *E*, but by initiating a process that could have caused a different event instead. In other words, holding the past, the natural laws, and the character of the agent[10] constant, the agent could have not-*E*-ed instead of *E*-ed. The

[10] Of course, the past and the laws of nature cause the character of the agent to be a certain way (either deterministically or indeterministically), but it's helpful to emphasize that we hold the agent constant.

chance to do some not-*E* might not be as high as the chance to *E*, but it's still physically possible. Thus the account serves the libertarian purpose. The agent had 'strong' alternatives to *E*-ing: she could have done some not-*E* instead, given the past and the natural laws.

Unfortunately, as I mentioned, the account doesn't satisfy (1'), so it doesn't allow for moral responsibility. Remember that (1') requires the agent to cause *E* (instead of *A)*, where *A* is one of the agent's alternatives to *E*. The agent could have acted the same way (initiated the same causal process) and *A* could have occurred. Nothing the agent did made the difference. Instead, the chance involved in the indeterministic process made the difference. Of course, the agent could have chosen some other indeterministic process that had a greater chance of causing *A* and a lesser chance of causing *E*, so the agent might still be somewhat responsible for *E*'s occurring. However, she is certainly less responsible than if she had initiated a process that *always* caused *E* and *never* caused *A*, which is to say, if she had caused *E* deterministically.

Of course, the better-known accounts recognize and try to get around that problem, but they can't. I don't expect the reader to take that on faith, so I'll briefly run through the major event-causal libertarian accounts and explain how they fail.

III. Three Event-Causal Libertarian Positions: Ekstrom, Kane and Mele

I start with Ekstrom's view because it is the most similar to the simple account just given. Ekstrom focuses heavily on the self and self-formation in her account of responsibility. In Chapter II I discuss Ekstrom's account of the self in more detail, but I only need the broad strokes here. Suffice it to say that she thinks the self is composed of

certain beliefs and desires that cohere (provide mutual justification), as well as a deliberative faculty[11]. Those special beliefs and desires are called acceptances and preferences, respectively. When an agent deliberates about an action, she considers her desires, preferences and acceptances that have some bearing on the situation. Based on that deliberation, the agent forms a preference about what to do in that situation, which causes, "in a normal sort of way[12]," an intention to act, which in turn causes an action. The desires, preferences and acceptances that the agent considers during her deliberation constitute her reasons for acting, and the preferences and acceptances account for her (self) causing the resulting preference.

Ekstrom injects indeterminism into her account between the reasons (or the self) and the new preference that leads deterministically to action. The reasons that occur to the agent indeterministically cause the preference that leads to action, so while those reasons are causal antecedents of the action, they do not necessarily (deterministically) cause it. The same agent with the same reasons in the same situation could form different preferences and different intentions, and thus perform different actions. Ekstrom's account therefore captures the libertarian requirement of 'strong' alternatives.

It might seem as if Ekstrom's account has the same straightforward flaw as the simple event-causal libertarian account. The problem, remember, is that since the very same agent, in the very same circumstances, could cause different actions, the agent cannot have caused one action *instead of* its alternatives. On Ekstrom's account, the same set of desires, preferences and acceptances in a given situation can result in different actions, and have no way of causing one action to happen *instead of* another, so it seems

[11] Laura Waddell Ekstrom, "Indeterminist Free Action," in *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001), 144.
[12] Ibid 145.

she cannot escape the problem. The agent doesn't make the difference between *E*'s happening and *A*'s happening.

However, as mentioned above, preferences actually count as part of an agent's self, and Ekstrom places the indeterminacy during preference formation. Thus, if an agent causes *E*, a part of her self does, in fact, cause *E instead of A*. During deliberation, though, that preference hasn't been formed yet, so nothing determines which event the agent will cause, so the agent could cause either one. Agents during deliberation, therefore, have 'strong' alternatives, but agents at the time of action don't. Thus Ekstrom's account allows the agent (at the time of action) to cause *E instead of A*, while preserving 'strong' alternatives[13].

I don't think that wrinkle can do all the work that libertarians need it to, though. It simply pushes the problem back a step or, more specifically, pushes it back a preference. The problem might be solved with respect to the action, but it isn't solved with respect to that preference and, by extension, the agent's self. The agent (prior to deliberation) can cause the new preference to form, but it cannot cause that preference *instead of* another one. Ekstrom acknowledges that problem, but chooses to ignore it, simply stipulating that agents don't have to be responsible for their preference formation.[14] Since I'm inclined to agree that agents don't need to be responsible for their selves, I can't really attack that stipulation, but other libertarians would probably find fault with it.

However, more importantly, the temporal difference between the agent's having 'strong' alternatives and the agent's causing *E instead of A* means that the agent can't

[13] Esktrom never mentions this facet of her account, though, so I'm not sure she intended it to work that way or not. Even if she didn't, though, it's an interesting move.
[14] Ibid 150. Ekstrom speaks of free acts, rather than morally responsible ones, but in this context I take them as similar enough as makes no difference.

actually satisfy (2'). Remember, (2') states that, if *S* is morally responsible for *E* at *t*, then at t, either *S* could have done something other than *E*, or is morally responsible for the fact that it is impossible for her to do something other than *E*. As mentioned, at *t* the agent cannot have done something other than *E* (in the strong sense), because the agent has formed a preference to do *E*, and that preference deterministically causes *E*. The only available option, then, is the second part of the disjunction. *S* must be morally responsible for the fact that she cannot do otherwise at *t*. She is causally responsible for that fact: her deliberation (indeterministically) caused the new preference to form, which made it impossible for her to do otherwise. However, she isn't morally responsible, by Ekstrom's stipulation. Ekstrom's account cannot, therefore, serve the libertarian's purpose: it allows for 'strong' alternate possibilities, but doesn't satisfy (2').

Robert Kane takes a different approach to the problem, though he places the indeterminism in a similar place. He relies on the notion of parallel processing in the brain, where the same agent tries to do two mutually exclusive things at once. His example is a businesswoman who, on her way to an important meeting, sees someone getting mugged.[15] She has two conflicting sets of reasons for acting, one supporting her continuing on to her meeting, the other supporting her turning back to help the person being mugged. Her brain, says Kane, is built in such a way that it tries to cause her to act on both of those sets of reasons at once, which of course she can't do. Kane calls each attempt to act an 'effort.'[16] The resulting internal conflict somehow stirs up the quantum indeterminacies in her brain, "temporarily screen[ing] off complete determination by the

[15] Robert Kane, "Responsibility, Luck and Chance: Reflections on Free Will and Determinism," in *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001), 164.
[16] Ibid 170.

past."[17] This indeterminacy eventually results in one of her efforts winning out, so that she acts on the reasons associated with it and not the other reasons. Kane calls the actions that result from multiple, competing efforts 'self-forming actions,'[18] or SFAs.

I should also mention that Kane doesn't think we go through this process when faced with just any decision, such as what to have for breakfast or whether to lock the front door at night. In cases like those, our existing characters and motives don't contain any self-conflict, so there's no reason to struggle against ourselves in making the decision. SFAs help create those existing characters and motivations, I guess by eliminating internal conflicts. Whatever outcome the SFA has, the agent can identify with it, can 'endorse'[19] it, and incorporates the decision as part of her established character.

Kane isn't totally clear about the order of events in SFAs, but I think the following is the most charitable interpretation: The conflict of efforts indeterministically causes an intention associated with one of the efforts, and that intention deterministically causes an action. Some of my argument below relies on that interpretation, and since I argue that the account doesn't work, the reader might think that I've deliberately chosen an untenable interpretation. However, the only other likely interpretation is as follows: each effort of will includes an intention to act, and the conflict of efforts directly and indeterministically causes the action. That means that at least some of the agent's intentions to take a particular action (partially) cause the agent to do something other than that action: the conflict of efforts (and intentions) caused the event that resulted in action. Further, which intention serves its proper purpose and which doesn't is out of the agent's

[17] Ibid 164.
[18] Ibid 163.
[19] Ibid 171.

control. That interpretation, therefore, makes Kane as vulnerable as the straw man I considered above, so I use the former interpretation.

How is Kane's parallel-processing picture supposed to solve the libertarians' problem? Well, consider the businesswoman example. Suppose the businesswoman turns back to help the person getting mugged. In that case, her action is backed by a set of reasons that caused her to turn back *instead of* continuing to her meeting. For example, she knew that turning back was the right thing to do, morally speaking, and that continuing to her meeting was the wrong thing to do. The event indeterministically caused by her efforts led to an intention that caused her to turn back. So Kane's account seems able to capture (1). However, at the time of deliberation, the businesswoman could have continued on to the meeting instead. That fits the second piece of the disjunction in (2'), since that event led to the intention that deterministically caused her to turn back, so the account fits (2') as well. Finally, at the time of deliberation, the businesswoman had the 'strong' alternative of continuing on, so Kane's account satisfies the libertarians' requirement.

That's Kane's argument, anyway. I'm not so sure it works. The problem starts, I think, with Kane's reliance on parallel processing within a single agent. It seems strange at best that a single agent could be literally trying to do two things at once, so that two "factions" within the person battle for control. If an agent is so divided as to have to defeat itself in making a decision, I have to question whether we can consider the agent's self a single entity. It makes more sense to see the agent as someone with a split personality. That isn't to say that we never experience internal conflict or turmoil. I just

think that such turmoil is best understood as competing desires, or of the agent against "alien" desires, rather than as the agent literally trying to do two things at once.

That's only a side issue, though. The main problem is that, as with Ekstrom, Kane only succeeds in moving the problem back a step. He's managed to give an account in which the agent causes *E instead of A*, yet has 'strong' alternatives to doing so. Unfortunately, nothing about the agent can account for why her *reasons* for causing *E instead of A* won out over her reasons for causing *A instead of E*. Since the reasons that win cause an intention that determinately cause *E*, we might consider reasons-causing-intention-causing-action as one complex event that follows the indeterministic struggle within the agent. Let's call that event *F*. We can also consider the other possibility at the time of deliberation, that the other set of reasons win and cause an intention that causes *A*, as a single complex event, *B*. The problem is that the agent cannot cause *F instead of B*; in other words, the agent cannot be responsible for the *E*-causing reasons' beating out the *A*-causing reasons. Since the agent isn't responsible for that, she isn't responsible for the fact that it's impossible for her to do otherwise than *E* at *t*. Consequently, Kane's account fails to satisfy (2'), despite initial appearances. Kane therefore fails to offer the libertarians a viable account.

Mele, recognizing that placing the indeterminism between intention and action won't work, and that trying to place it between reason and intention looks equally doomed, tries a different tack. Though we need to have control over our intentions and actions, we generally don't think of ourselves as needing control over which reasons occur to us. Mele takes advantage of this and injects the indeterminism into the process by which reasons occur to us. The causal story would therefore go something like this: *S*

deliberates about a choice. That deliberation includes reasons occurring to *S*, and they do so indeterminately: given the past and the laws of nature, different reasons could have occurred to *S* than those that in fact did. Once the reasons occur to *S*, her deliberation continues with her weighing the reasons and forming an intention based on them. The process of the reasons causing her to form an intention happens deterministically. That intention then leads deterministically to an action.

In justifying his account, Mele appeals to the motivation that I considered and rejected earlier. He says that libertarians may just value independence from the past, and that his account supplies that independence. He uses an imaginary libertarian, Wilma, as an example of someone with that value:

> [Wilma] reports that the thought of her actions as links in a long deterministic causal chain is somewhat deflating and that the truth of determinism is inconsistent with her life's being as important and meaningful as she hopes it is…Wilma values…a measure of independence from the *past*. She values, she says, a kind of independent agency that includes the power to make a special kind of contribution to some of her actions and to her world…The kind of agency she hopes for, Wilma says, would render her decisions and actions more personally meaningful from the perspective of her own system of values than they would otherwise be.[20]

I can't argue against a value like that, but I can argue that the value doesn't seem to have anything to do with moral responsibility. The best way to see that is by analogy. Imagine that you and your friend go to a video arcade in which there are two seemingly identical Tetris machines. In fact, they are identical, save in one respect. One machine uses a regular, deterministic computer program to determine the order in which the game

[20] Alfred Mele, *Free Will and Luck* (New York: Oxford University Press, 2006), 100. Interestingly, Mele doesn't have even Wilma believe that indeterminism is necessary for moral responsibility, but that's because he's agnostic about whether moral responsibility is compatible with determinism, indeterminism, both, or neither.

pieces appear. The program uses a "random seed" (based on the date and time) and a complex algorithm such that the pieces appear in an unpredictable order, but that order isn't truly random. Given the same random seed, the pieces would appear in the same order. The other uses an experimental program based on quantum physics, which makes the order of the pieces truly random. You certainly wouldn't be able to tell the difference between the two machines, though. You play a series of ten games on the normal machine, and your friend plays ten games on the indeterministic machine. Is your friend more responsible for his score than you are for yours?

Of course not. Your friend simply took the pieces as they came and tried to do the best he could with them, as you did. The fact that the order of his pieces was independent of the past doesn't make him more 'deeply' responsible. It doesn't make any difference at all. The same applies to Mele's account. The fact that the agent's reasons occur indeterministically doesn't make her more 'deeply' responsible. She simply took the reasons that occurred to her and tried to do the best she could with them, exactly as she would if they occurred deterministically.

That's the problem with Mele's account: it gives the agent 'strong' alternatives, but not the kind of alternatives in which libertarians are interested. Since the agent has no control over the reasons that occur to her[21], we can consider those reasons an external factor, just part of the situation the agent faces. We don't create reasons, or identify with reasons, unless we're speaking very loosely; one can "identify" with a political ideology, for instance. Reasons simply *occur* to us, as restatements of environmental pressures.

[21] Mele's full account gives the agent some form of control over those reasons: he thinks that a person's past decisions affect the probability of reasons' occurring. That doesn't amount to the control necessary for moral responsibility, though; it doesn't satisfy (1b), for one thing. It also doesn't make the reasons any more internal to the agent.

Mele's account only offers alternatives of which *situation* the agent faces: is this a situation where I think of *x* reasons to act, or *y* reasons? My character makes no difference in the outcome, just an indeterministic process that occurs before I even get involved.

Ekstrom, Mele and Kane all fail to overcome the libertarians' major problem. The irony is that the requirement they cannot meet is a result of the very intuition they think is so important. They focus on the importance of the agent's having alternatives for moral responsibility, but their insistence on 'strong' alternatives cuts the agent off from controlling any option at all. Based on everything above, we can tentatively conclude that requiring 'strong' alternatives means requiring inconsistent conditions for moral responsibility, and so makes moral responsibility impossible.

How should we weaken the requirement, then? One option is to throw it away completely. Frankfurt famously makes that move in his attack on the Principle of Alternate Possibilities.[22] He convinced many philosophers, John Martin Fischer included, that moral responsibility doesn't require the ability to do otherwise at all. All it takes is control and lack of coercion, on that view. I think totally throwing out (2) is too extreme, though. As mentioned, it's based on the fact that we praise and blame based on the contrast between alternatives, and I haven't seen a good reason to give up that practice. Besides, I don't think the "Frankfurt-style cases" really show what Frankfurt, Fischer and others think they show.

[22] See Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66 (1969): 829-839.

IV. Against Frankfurt-Style Cases

Although most people familiar with the free will debate are also familiar with Frankfurt-style cases, I need to run over it briefly before I continue. For one thing, it can't hurt to have the case ready at hand. For another, I will also discuss a variation, proposed by Mele, on the classic case, and it may help to compare the two directly. Frankfurt's case runs as follows: Black, a capable neurosurgeon, wants Jones to perform a certain action. To that end, he wires Jones' brain such that, if Jones decides not to perform the action, Black can force him to perform it. But Jones happens to decide to perform the action, so Black never has to coerce him into doing it. Is Jones morally responsible for performing the action? Intuitively, we would say, yes. But Jones doesn't have any alternatives to performing the action: if he hadn't decided to do it on his own, Black would have forced him to. Because of that, Frankfurt concludes that moral responsibility doesn't require alternate possibilities.

Some libertarians have pointed out that Frankfurt doesn't mention whether Jones lives in a deterministic or indeterministic world. If it's deterministic, they say, then Frankfurt's conclusion doesn't matter to them: they don't think Jones was free in the first place. If it's indeterministic, on the other hand, then Black's coercive mechanism won't work properly. By definition, one cannot predict with certainty the outcome of an indeterministic event, and Jones's choice is such an event. Therefore, if Black forces Jones to decide to perform the action, one of two things must happen. Either Black jumps the gun and forces Jones to decide before he does so on his own, in which case Jones never makes an autonomous decision and so isn't responsible, or Jones decides on his

own to not take the action and Black forces him to change his mind, in which case Black simply coerces Jones and Jones isn't responsible.

I don't have this route available, though, because I'm not a libertarian. Fortunately, there's another move that works just as well. Let's assume Jones lives in a deterministic world. As in the indeterministic case, Black can't wait until after the decision to activate his coercive mechanism. In this case, though, Black might have a reliable indicator of Jones's future decision. Assume, for instance, that if and only if Jones blushes at t_1 then he will decide to perform the action Black wants at t_2. Black could wait until t_1, and if Jones blushes he can refrain from activating his coercive mechanism. Jones will therefore decide to perform the action on his own. If Jones doesn't blush, he can activate his mechanism, and so force Jones to decide to perform the action he wants. Thus, if Jones in fact blushes as t_1 and performs Black's desired action, then he seems morally responsible and yet could not have done otherwise.

But wait. Imagine that the world is deterministic (it isn't hard to do!), and that you have limited prescience. Specifically, you can see murders before they happen. Unlike in *Minority Report*, if you "pre-see" a murder, it *always* happens. Now imagine that you "pre-see" Bert murdering Ernie. It certainly seems right that we should blame Bert *now* for murdering Ernie *later*[23]. That is, Bert seems morally responsible for the murder, because we know it's going to happen. Your vision of the murder is equivalent, at least for the purposes of assigning moral responsibility, to the murder itself.

Now consider Frankfurt's case. If Jones lives in a deterministic universe, and his blushing at t_1 entails his deciding to perform the action at t_2, then if he blushes, we know

[23] That is, we should blame him if we accept a compatibilist understanding of moral responsibility. I'm talking about a compatibilist response to Frankfurt, so that's appropriate.

at t_1 that he'll decide to perform the action at t_2 (assuming no manipulation). Just like your hypothetical prescience of the murder was equivalent to the murder, his blushing at t_1 is equivalent (at least for the purposes of praise and blame) to his deciding to perform the action at t_2. And if his not blushing at t_1 entails his deciding not to perform the action, then his not blushing is equivalent to his deciding not to perform the action. In other words, at t_1 we can assign the praise and blame for his decision at t_2 (when Jones blushes or doesn't). *Blushing* becomes the morally relevant act, because it's a perfect indicator of what Jones would have done, *sans* manipulation. That means that, if Black waits until Jones blushes (or doesn't), he's still effectively waiting until after the decision to activate his coercive mechanism. The Frankfurt case just looks like straight-up coercion, in which case Jones is obviously not morally responsible. We intuitively say he is because Frankfurt offers an impossible scenario, in which Black activates his mechanism only if Jones would decide against his wishes, but can't know if Jones will or not (unless Jones has already effectively decided).

Some philosophers have attempted to reformulate Frankfurt's case so that it gets around the problems I just discussed, Mele among them. His case is framed as a response to the libertarian counter, but applies to the compatibilist one I've suggested, because it avoids the coercive mechanism's dependence on some prior sign of a decision. I don't think it works, though. Here's Mele's case:

> …Black initiates a certain internally deterministic process *P* in Bob's brain at t_1 with the intention of thereby causing Bob to decide at t_2 (an hour later, say) to steal Ann's car. The process, which is screened off from Bob's consciousness, will culminate in Bob's deciding at t_2 to steal Ann's car unless he decides on his own at t_2 to steal it or is incapable at t_2 of making a decision (because, for example, he is dead by t_2). The process is in no way sensitive to any "sign" of

> what Bob will decide. As it happens, Bob decides on his own to steal the car, on the basis of his own indeterministic deliberation about whether to steal it, and his decision is not deterministically caused. But, if he had not just then decided on his own to steal it, *P* would have issued, at t_2, in his deciding to steal it.[24]

The questionable premise in this case is that *P* can be set to activate at t_2, contingent on Bob's decision at t_2. Can *P* activate at t_2 if Bob chooses not to steal the car at t_2, and not activate if he does choose to steal the car at t_2, without interfering with Bob's causal process? I don't think so. My argument gets a little technical, so I'll use some variables to represent the various actions. Let X_1 stand for Bob deciding to steal the car on his own. Let X_2 stand for Bob deciding to steal the car as a result of *P*. Let Y_1 stand for Bob's deciding not to steal the car[25]. Note that X_1 and X_2 are distinct, albeit fine-grained, events. In the case Mele uses, X_1 is some sort of indeterministic process, maybe something like Kane's efforts, while X_2 is a deterministic process initiated by *P*.

Now consider the possibilities. Either X_1 or X_2 can happen, but not Y_1. X_1's occurring rules out both other events' occurrence, and X_2 rules out Y_1's occurrence. In other words, either Bob will decide to steal the car on his own (X_1) or he won't. If he does, that rules out *P*'s causing him to steal the car (X_2), by stipulation, as well as his deciding not to steal the car (Y_1), by the principle of non-contradiction. If he doesn't, then in the absence of manipulation he would have decided not to steal the car (Y_1), by the law of excluded middle. However, if Bob doesn't steal the car on his own he *does* get manipulated, through *P*, so X_2 happens and Y_1 doesn't.

---

[24] Mele 88.

[25] I don't include one possibility here: Bob could have not decided anything at all by t_2. It has the same consequences for Bob's responsibility, though (he didn't decide to steal the car on his own), so we can just lump it in with Y_1.

Suppose X_1 occurs. We can know after the fact that Bob freely decided to steal the car, and that he would have done so even if Black hadn't tinkered with his brain. In that light, it certainly seems clear that Bob is morally responsible for his decision. But the only other possibility was X_2, where Bob still decides to steal the car. It seems, therefore, that Bob is morally responsible without being able to do otherwise.

Consider X_2 a little more closely, though. What does X_2's occurrence tell us about Bob? It tells us that Bob didn't choose to steal the car on his own. If Black hadn't tinkered with Bob's brain, Bob's only two options were X_1 and Y_1. So X_2's occurrence tells us that, if it hadn't been for P, Y_1 would have occurred. If X_2 occurs, then, we should treat Bob as if Y_1 had occurred. Thus the moral difference between X_1 and X_2 is identical to the moral difference between X_1 and Y_1. Bob therefore has two morally relevant options: X_1 and X_2. Mele's case leaves Bob with alternatives, and therefore can't show that alternate possibilities are irrelevant to moral responsibility.

V. The Other Option

We can't just eliminate (2'), as Frankfurt thinks we can, but neither can we stick to 'strong' alternatives, as libertarians want. Can we find a middle ground? Some compatibilists think so. They think we should understand (2') in such a way that it doesn't require S to be literally able to do other than E. There are a few common readings, but I don't think any of them quite get the job done. On one, we should interpret 'could do something other than E' to mean 'would have done something other than E, had he chosen to do something other than E.'[26] There's an obvious problem with this interpretation, though. What if the E we are discussing is a decision itself, say, to take

[26] Roderick Chisolm, "Human Freedom and the Self," in *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001), 128.

action *F*? We often are concerned with responsibility for decisions, as in the Frankfurt-style cases. In that case, (2') would read

(2'$_{ch}$) At *t, S* would have done something other than decide to *F*, if *S* had decided to do something other than decide to *F*, or is morally responsible for the fact that it is impossible for her to do something other than *E*.

Aside from being very unwieldy, (2'$_{ch}$) has the agent deciding to decide to *F*. We now have a second-level decision. Surely, if *S* is responsible for her decision to *F*, she must be responsible for her decision to decide to *F*. But if she is, then she must have been able to decide not to decide to decide to *F*, and now we're clearly caught in an infinite regress.

Another suggestion: maybe we should read 'could have done something other than *E*' as 'would have done something other than *E*, had the situation been different.' In other words, had different reasons occurred to *S*, or if she had been in a better mood at the time, or if that bird hadn't narrowly missed her head just before she faced the decision, she would have done *A* instead. The problem with this move, I think, is that it focuses on the wrong aspect of the decision-making process. Of course the agent will react differently to different situational stimuli. So will a computer. So will a dog. So will bacteria. Who cares that, if John Wilkes Booth had been struck in the head on his way to Ford Theatre, he would have calmly taken his seat and watched the show with the rest of the audience? That's not the kind of alternate possibility in which we're interested. We want to talk about differences the *agent* makes, not the circumstances. I haven't seen a compatibilist formulation that properly focuses on that difference. So here's one. If *S* is morally responsibly for *E*'s happening at *t*, then

(2'$_{DA}$) A different agent, in the same situation as *S*, would have done something other than *E* at *t*, or *S* is morally responsible for the fact that every agent would have done *E* at *t*.

In other words, *S* caused *E* (*instead of A)*, but would have caused *A (instead of E)*, if *S* had been a different person. That formulation places the emphasis where it belongs: the difference between the *agents* makes the difference between the actions. I imagine that (2'$_{DA}$) might provoke a few objections, though. I'll deal with three below.

First, it might seem strange that *S*'s being responsible is tied to someone else's actions (or potential actions). That's a common objection against David Lewis's account of free will using counterpart theory. I agree that it's strange to think that my counterparts' actions are what make me responsible, but those are supposed to be real actions (just not actual ones), and that's not what I'm suggesting. I'm saying that, when we ask if *S* is morally responsible for *E*, we should ask, "Would anyone else have done any differently?" We often ask ourselves such questions, and so tie one person's responsibility to others' potential actions, when we're actually considering people's behavior; why not let that practice inform theory?

Second, what about cases where agents have very different causal powers? For example, a gun to my back would affect my decision-making very differently than a gun to Superman's back would Superman's decision-making. Surely Superman's power to uppercut his would-be coercer into orbit shouldn't have any bearing on my moral responsibility for caving to the gunman's wishes. I agree that agents with different causal powers shouldn't count when we're determining if *S* is morally responsible. Similarly, we should only compare agents with the same rational powers, information about the

situation, and so on. However, I define the agent in such a way that rules out the problem. I discuss this in much greater length in Chapter II, but suffice it to say that I distinguish the agent from many of the mental and physical states that we often attribute to agents. All that should count for (2'$_{DA}$) is the agent's 'true self,' which I identify as something called the 'preference set' in that chapter. Preference sets only have one function, which is to take as input beliefs and desires and create intentions from them. All other causal powers that we normally attribute to an agent actually belong to the agent's body or brain, and we hold the body and brain constant for (2'$_{DA}$). So Superman's decision is relevant, but only if we frame the question as follows: "What would Superman do if he were in my situation, in my body, and had my brain [except for the part concerning the preference set; again, see Chapter II]?"

Third, many philosophers think that when we deliberate, it seems to us as if we're choosing from genuine alternatives. When I'm thinking about whether to have a turkey sandwich or a salad for lunch, and I choose the salad, it seems to me as if I could have chosen the sandwich, and not in the sense that another person would have chosen it in my situation. If the meaning of 'could' in rational deliberation means something other than my understanding of 'could,' why should we think my understanding is relevant to rational deliberation and, by extension, moral responsibility? Well, I can say for certain that at least some people see their own deliberation as a process that doesn't involve 'strong' alternatives. I, for one, see myself that way. Again, I'll discuss the process more in the next chapter, but I imagine my decision goes something like this: I first recognize my options, which are dictated by external conditions; for example, I know I can have a turkey sandwich or a salad, because I have the materials for both in my refrigerator, but I

cannot have a steak, because there isn't one in my refrigerator and I don't have time to buy one. I then consider my reasons for choosing each one; for example, that the salad has fewer calories and more Vitamin A, but the sandwich has more protein and tastes better. I rank those reasons in order of importance to me, and I act on the set that I find more important overall. Let's say I choose the sandwich.

My decision about lunch is an indeterminate (or maybe incomplete) event before I get involved: the reasons for picking each meal can't, by themselves, determine which thing I'll eat. However, my involvement makes the event determinate: I *determine* which meal I want to have. I make one "possibility" actual, and in that sense choose from among real alternatives, even if *I* would always make the same choice in that situation.

(2'$_{DA}$) captures that kind of alternative, and looks consistent with (1'): in the example above, I caused myself to eat the sandwich *instead of* the salad, intended to eat the sandwich *instead of* the salad, and did so in the manner I intended. If we accept (1') and (2'$_{DA}$) as adequate representations of their underlying intuitions, then, we have hope for a consistent set of necessary conditions for moral responsibility. We're not out of the woods yet, though. I still have one more intuition to handle.

VI. The Third Intuition

The last intuition is not as universal as the other two, in my experience, but has played a significant enough role in the debate that it warrants discussion. It's usually referred to as 'ultimate control' or 'original control,' as goes like this: If *S* is morally responsible for *E*'s happening at *t*, then

(3) *S* is morally responsible for the relevant characteristics of her self at *t*.

By 'self' I mean the agent's "true self," as mentioned above and discussed in the next chapter. The motivation behind the intuition seems pretty simple. We tend to think of agents as having characteristics, or dispositions to act in certain ways, and those traits or characteristics (unsurprisingly) inform the agents' actions. If the agent acts because of and in accordance with those traits, it seems like the agents need to be responsible for having them. For instance, we think that Faust made his deal with Mephistopheles because he wanted knowledge and power, but also because he was a *bad person*; he was greedy and vain and willing to deal with the Devil to get what he wanted. But what if Faust was just born that way? What if he had no control over whether or not he was greedy and vain? What if he were simply acting out his role with the programming he received? Why should we blame him for that? He couldn't do anything else but act as his character, his self, dictated, even though he faced meaningful options in applying that self. Therefore, in order to be responsible for his actions, he must be responsible for his self. If we accept that, though, we have a serious problem: (3) leads directly to an infinite regress.

(3) is a reflexive condition: it's a condition for moral responsibility that refers to the conditions for moral responsibility. Assume that *S* caused *E* because she had some quality *Q*. Assume she had *Q* because she made an earlier decision to adopt *Q* as a trait, based on a quality she had at that time, Q_0. If we assume *S* is morally responsible for *E*, that means that

S was in control of *E*'s, instead of *A*'s, happening at *t*.
S could have done something other than *E* at *t* (understanding 'could' as discussed above).[27]
S was morally responsible for the characteristics of her self at *t*, which entails
    S was in control of her self having quality *Q* instead of quality *R*.
    S could have done something other than cause herself to have *Q*.
    S was morally responsible for the characteristics of her self (as it existed when she came to have *Q*), which entails
        S was in control of her self having quality Q_0 instead of quality R_0.
        …etc.

It's quite obvious that the regress cannot go on infinitely, because agents are not infinite or eternal beings. Humans, for instance, don't seem to become agents for several years after birth.

So consider that first instant of being an agent, whenever it is, for *S*. There are only two possibilities concerning her self. First, *S* might have no personality traits at all, and gotten them all ex nihilo, either as soon as she became an agent or over time, as in Kane's account. Even if that were true (and empirically it doesn't seem to be), it wouldn't help ground moral responsibility. For what mechanism does *S* use to create her traits? Upon what criterion does she choose one set of traits over another? If it's some internal criterion, it sounds an awful lot like an innate trait, and *S* clearly doesn't have control over which traits she has innately. If it's an external criterion, how did she come to use it? She must have chosen it, but then how did she choose? She must have used some criterion, and so we're stuck in a regress. If there's no criterion at all, and the agent simply leaves it up to chance, she seems no more responsible than if she had just started with some random set of traits. That's the second possibility, of course; she might have

[27] Note that the regress doesn't depend on the content of (1) and (2). I've included those conditions for completeness.

started out with some traits for which she was not responsible. Neither possibility allows the agent to satisfy (3), so (3) seems to be an impossible condition.

I should mention that Mele attempts to get around the problem in a very peculiar way: he claims that, at the beginning of agency, we should relax the conditions for moral responsibility[28]. At that time in a person's life, he argues, the decisions they make aren't as important. They choose between, for instance, stealing a cookie or not, yelling at their sibling or not, brushing their teeth or not. Since the choices are so trivial, it should be easier to be morally responsible for them. So maybe "little agents"[29] don't need to be responsible for themselves, but as they develop selves, they have to start being responsible for them.

Mele's argument has a glaring flaw, though. The early decisions in the agent's life become the basis for her developing personality, on his view.[30] The agent's self, and by extension all her actions in her entire life, rest on the decisions she makes at the start. If we look at a fully-formed agent, according to Mele, we can trace each of their traits back to one of her early decisions. A life of crime sprang from a seed planted with a fistful of stolen cookies, and a perfect saint began as an unerring teeth-brusher. Given the implications those decisions have for an agent's later character, they are by far the most important she'll ever make, not the least. If anything, we should have stronger requirements for holding her responsible for them, not weaker.

Though most philosophers recognize the problem, not all are willing to concede the point. It seems to me that many of the agent-causalist libertarians are driven to their

[28] Mele 130.
[29] Ibid 129.
[30] He thinks they raise the probabilities of taking certain actions and thus forming new traits, not that they determine new traits, but whatever the mechanism, they still account for the agential contribution to acting and forming new traits.

position by the problem with (3). They hope that, by introducing agent-causation (a kind of substance causation), they can avoid dealing with the event-causal paradox I've described. Unfortunately, I don't think the move helps. Suppose the agent is free from the bounds of the past and the laws of nature. It's still subject to the laws of logic. Any time it makes a decision (say, to act in accordance with one set of reasons over another), it must do so either in accordance with some principle or randomly. If the agent-substance does so randomly, it seems strange to blame it for the decision. If it uses some criterion, though, it's placed back in the dilemma that faced our event-causalists: Is that criterion internal or external to the agent? If it's internal to the agent, it seems the agent needs to be responsible for having it…and so on.

Given that (3) leads immediately to a regress, it doesn't seem likely to me that we will save it. It's pretty difficult to just throw out an intuition, but I don't think we have any choice. If we want to maintain that moral responsibility is logically possible, we have to throw out intuitions that require impossible conditions. As consolation, though, I think the reason (3) seems necessary to many people, is that they think of an agent's self, its characteristics or traits, as somehow separate from the agent. I don't have an argument for this, exactly, but it seems almost incoherent to me that we think agents should have control over *being a certain kind of agent*. People separate agents from their traits because, in everyday language, we often identify the agent with things that we know aren't identical to the agent. For instance, we say that Mary *is* angry, or Matt *is* tall. We have good reasons not to identify agents with their emotions or bodies. But we also say that a person *is* evil, or pleasant, or timid, and we *do* have reason to identify them with those kinds of traits. I explain why in the next chapter.

VII. Conclusion

In this chapter I have examined the three intuitions concerning moral responsibility and made explicit the requirements each state. I merely clarified (1), but I substantially altered (2) and threw out (3) altogether. In their post-reflective state, the intuitions are as follows: In order for *S* to be morally responsible for *E* at *t*,

(1a') *S* caused (*E* instead of *A*) to happen.

(1b') *S* intended to cause (*E* instead of *A*) to happen.

(1c') *S* intended to cause (*E* instead of *A*) to happen in the way, and at the time, she *in fact* caused (*E* instead of *A*) to happen.

(2'$_{DA}$) A different agent, in the same situation as *S*, would have done *A* instead of *E* at *t*, or *S* is morally responsible for the fact that every agent would have done *E* at *t*.

Those conditions are compatible with determinism, and incompatible with indeterminism, as I have discussed above. It seems to me, then, that compatibilism is the most reasonable position for a philosopher unwilling to give up on the notion that moral responsibility is logically possible. The question about whether *we* are in fact morally responsible still remains, of course.

The answer depends on a number of things. First, we need determinism to turn out true[31]. Second, we need to be the kinds of things that could possibly be morally responsible, which is to say, we need to be agents. Third, we need to interact with the world in such a way that we can exercise that moral responsibility, which is to say, we

[31] More specifically, we need *local* determinism to turn out true: we need the events concerning action to be deterministic. Fortunately, those look like "macro-level" events, so we can dodge the bullet of quantum physics (which leaves a spooky wave pattern on anything it hits).

need to act autonomously (that is, free from coercion or manipulation). Sadly, there's no consensus on what it takes to meet the latter two conditions. I tackle that problem next.

Chapter II: Autonomy and the Self

In the last chapter I analyzed three intuitions that established necessary conditions for morally responsible action. As I mentioned in the introduction, however, I think the vagueness in the free will debate extends past those three intuitions. In particular, participants in the debate rarely discuss what it means to act, or even to be an agent[32]. That's a problem, I think; after all, the debate concerns *agents*, and whether or not they are morally responsible for their *actions*.

Tautologically, an agent is simply something that acts. Many philosophers seem content to leave the definition at something as unhelpful as that, assuming that people intuitively understand what they mean by "agent". This has led to a fair amount of confusion, including the general acceptance of a theory of action that I believe makes autonomous action, and with it morally responsible action, impossible.

In this chapter I attempt to provide that clear conception of the agent. I offer an account of the self that coheres with our intuitions about autonomy, as well as with the three (revised) intuitions in the previous chapter. My position centers on a mental phenomenon which I will call the preference set, and which I describe in Section III, below. A review of the standard story of action, as well as a few prominent models of the self, will help put my position in context.

I. The Standard Story of Action and its Problems

Most philosophers recognize the existence of at least two types of mental state: beliefs and desires[33]. A belief is a mental state in which one holds that a proposition is

[32] More precisely, they rarely discuss action and agency in the context of the free will debate.
[33] Eliminative materialists don't believe in beliefs or desires, but I'll leave that debate for another day.

true. Desires are positive attitudes towards a state of affairs[34]. Both beliefs and desires are impermanent mental states: they can and do change quite often. Many philosophers also recognize the intention as a separate kind of mental state, one which, under normal circumstances, leads to an action. An intention is usually treated as a kind of bridge between mental states and action.

How exactly are beliefs, desires and intentions supposed to contribute to a person's action? In his important essay, "What Happens when Someone Acts?" J. David Velleman describes the "standard story of human action":

> There is something that the agent wants, and there is an action that he believes conducive to its attainment. His desire for the end, and his belief in the action as a means, justify taking the action, and they jointly cause an intention to take it, which in turn causes the corresponding movements of the agent's body. Provided that these causal processes take their normal course, the agent's movements consummate an action, and his motivating desire and belief constitute his reasons for acting.[35]

Note that, in this standard story, there does not seem to be a whole lot of room for the agent herself to affect the goings on. Unless the agent is identical to the desire, or the belief, or the intention, or a combination thereof, she is not involved in the process at all. That poses a clear problem for those who want to hold that agents act autonomously: if the agent is not involved in the causal process leading to an action, she probably is not acting autonomously (it is hard to say how she is acting at all). There are two standard solutions to the problem, but as Velleman argues,[36] those solutions don't work.

---

[34] Philosophers don't agree on the kind of attitude desires are (simply dispositions to act? Dispositions to take pleasure from a state of affairs? Beliefs that states of affairs are good?), but I don't want to get into that, either. My vague definition should suffice for this essay.
[35] Velleman 461.
[36] Ibid 462-5.

One of the proposals is to hold that, since the causal processes are occurring within the agent (since the agent is the "subject"[37] of the mental causal processes), she is involved in that way. However, it is important here to point out an ambiguity in the use of the word 'agent' which people often overlook. Sometimes we use 'agent' to refer to the sum of all of a person's mental structure and activity, and often a person's body as well. Sometimes, though, we refer to only a portion of a person's mental structure and activity. One might endorse different claims about which part of a person's mental life counts as the agent[38], but it is clear that we do not always identify the agent with everything that occurs in her mind or body.

For example, consider the case of Frankfurt's unwilling addict[39], who wishes she could refrain from taking a drug but is unable to stop herself from doing so. In this case the addict is the host of a causal process in which a desire to take the drug couples with a belief that doing *X* will result in taking the drug, and together they cause the formation of an intention to do *X*, which in turn causes the addict's body to do *X*. The agent feels horrified by and alienated from her desires and actions, and tries everything she can to stop herself from taking the drug, but fails. We might say that the agent "couldn't help" but take the drug or "was powerless" to refrain from taking the drug" In using such phrases we clearly distinguish between the agent and the mental and physical processes occurring in her mind and body. If the desire that led the agent to take the drug always counted as part of her, and that desire compelled her to take the drug, it wouldn't make sense to call her powerless. Thus we have a case in which the process described in the standard story of action occurs in an agent, but in which we would not identify the agent

[37] Ibid 463.
[38] Ekstrom offers an account in opposition to mine. I outline her position and discuss it later in this chapter.
[39] Frankfurt 1971, p. 83.

(in the narrow, "true self" sense) with the desire, belief and intention moving the agent (in the broad sense) to act. It seems clear to me that when we speak philosophically about agents we mean to use the narrower sense of the word, to talk about a person's "true self" and its influence on the person's behavior.

It is clear that the first proposed solution to the agential involvement problem fails precisely because of the above distinction. The solution works as long as we use the wider sense of agent: the causal process does indeed occur in the agent's mind. Unfortunately, the process's occurring there isn't sufficient to account for agential involvement, as the case of the unwilling addict shows. The agent in the narrower sense needs to be involved in the causal process leading to action, and the first proposed solution cannot provide that.

I should mention here that many proponents of the standard story disallow the above case and others like it, claiming that the cases are examples of abnormal behavior. They say that the standard story of action only applies to cases where the agent is not alienated from her beliefs, desires and intentions, or at least where they act without a desire so strong that it constitutes compulsion. That move drives Velleman to propose a different counterexample, in which he thinks a person behaves normally and yet is not involved in the action.[40] However, the example he gives includes a few hard-to-swallow assumptions about consciousness and responsibility. Further, I think his resort to the new example is unnecessary. The standard story's proponents' claim that the behavior exhibited by the unwilling addict is abnormal, while true, takes no bite out of the counterexample. It doesn't matter that such behavior occurs only infrequently, or that it involves compulsion. The unwilling addict still "hosts" the causal process that results in

[40] Velleman 464.

an action, and that hosting is supposed to constitute her involvement in a normal case. Since we don't think the unwilling addict acts autonomously, the first proposal fails.

The other proposed solution is to hold that the agent is indeed identical to the desire, belief and intention, or perhaps the conjunction of the three. The belief, desire and intention combined causing an action therefore amounts to the agent's causing an action. Justifying that move takes some work, though. It seems pretty clear that agents aren't just collections of desires, as the case of the unwilling addict shows. They aren't just some set of beliefs either: while agents must meet a certain epistemic requirement to be morally responsible, we don't blame a set of beliefs for killing MLK. Intentions don't fit the bill any better. We might blame someone for intending to steal the communion wine, but we don't literally blame that intention. It seems more intuitive to say that the agent is something different than desires, beliefs and intentions, and that it somehow interacts with them to cause events (like the motions of a person's body). Agents must be the kinds of thing that can choose between desires and beliefs in order to form an intention, and it does not seem at first glance that any desire, belief or intention could do that.

Besides, the agent can't be identical to all of her beliefs, desires and/or intentions. In the case of the unwilling addict, the agent doesn't identify (and we don't identify her) with her desire to take the drug, nor with her intention to take it. It also seems possible for a person to feel alienated from some set of beliefs; for example, someone raised in a racist environment might have an ingrained belief in a natural hierarchy of races, and feel terrible that she cannot rid herself of the belief. Anyone arguing for the second solution to the problem needs to give a more sophisticated account of how it works. Several

philosophers have tried, but I don't think their accounts can solve the problem. I deal with them in the next section.

II. Desire- and Belief-Centered Views of the Self

I will discuss three views that attempt to provide an explanation for how some set of beliefs or desires might stand in for the agent's involvement in an action[41]. The first is presented by Harry Frankfurt, the second by Gary Watson, and the third by Laura Ekstrom. I start with Frankfurt because his is the most widely-known position, usually called a hierarchical account of the self. Frankfurt begins by pointing out a distinction between types of desires that we have. We have first order desires, which are simply "desires to do or not do one thing or another."[42] However, we also have second-order desires, which are desires that take first order desires as their objects. So an agent might have a first order desire which takes the form, "I want to X" and a second-order desire of the form, "I want to want to X." There are special kinds of both first- and second-order desires. A first-order desire on which one acts is called one's will, and a second-order desire concerning which first-order desire one wants to be one's will is called a second-order volition. Suppose Susan has a first-order desire to go to the movies with her friends and a first-order desire to read Hegel for her seminar. She also has a second order desire that her first-order desire to read Hegel be her will. That second-order desire is a second-order volition, and her desire to read Hegel, since it in fact informs her action, is her will.

[41] Of course, more than three such views exist: Velleman offers one, for instance. However, these three cover a wide range of the various positions, and the objections I offer to these theories can, with a little work, be modified to apply to other, similar views.
[42] Frankfurt 1971, p. 78.

Frankfurt defines a person (which I think is supposed to be more-or-less identical to an autonomous agent) as a being which has second-order volitions.[43] This formulation seems to fit our intuitions pretty well. Imagine a being without any second-order volitions, a being Frankfurt calls a "wanton".[44] This being would go around satisfying his first order desires, whatever they were, without ever considering "the desirability of those desires themselves."[45] That is not to say that the being cannot act rationally: he could easily use reason to figure out the best way of satisfying his desires. However, he is certainly not a morally responsible person, for he does not even consider whether or not his desires are morally proper ones, let alone whether or not he wants to act on the ones that are morally proper. In fact, as mentioned above, Frankfurt seems to say that he is not a person at all. He has no self, just a set of desires and a body that goes about satisfying them.

On Frankfurt's model, a person acts freely (read: autonomously[46]) when the object of her second-order volition in is in fact her will. In the earlier examples, Susan acted freely when she read Hegel, while the poor drug addict did not when she succumbed to her addiction. So Frankfurt might solve the problem of involving the self in the standard story of action by saying that the self is represented by a special type of desire: the second-order volition. If that desire is one of the ones that are causally efficacious in the performance of an action, then the self is involved in the action and we

[43] Ibid. 82
[44] Ibid.
[45] Ibid.
[46] Freedom and autonomy are not generally synonymous, but in this case I believe they are being used so. Frankfurt's account is (or appears to be) compatible with determinism, and traditionally compatibilists believe that to act autonomously is to act freely (or at least as freely as we need to be in order to act responsibly). Incompatibilists disagree, of course. I don't want to conflate the two notions, but I want to use Frankfurt's original terminology, so if you are an incompatibilist, simply read "autonomously" any time I write "freely".

can say that the person acted. If it was not, as in the case of the drug addict, then we can say that the person did not act, but was "…hopelessly violated by [her] own desires."[47]

As many critics have pointed out, such a position faces a few serious problems. There are two problems most pertinent to our present discussion. First, Frankfurt offers no reason to support that a second-order volition is enough to ground autonomy. As Frankfurt admits, one could easily have several conflicting second-order desires, each corresponding to a different first-order desire[48]. For example, Susan has a second-order desire to read Hegel: she wants to want to read Hegel, because that would mean she's the kind of person her parents want her to be, and so on. However, she might also have a second-order desire to go to the movies: she wants to want to go to the movies, because she wants to be the kind of person her friends think of as fun, and so on. Why not appeal, then, to a third-order volition: maybe Susan wants to want to want to go to the movies, and wants that second-order desire to be volitional. Does Susan require a third order desire to act freely or not? If so, why not continue ad infinitum? If not, why not?

Second, the hierarchical aspect of the model bears almost none of the weight Frankfurt seems to think it does. An agent's involvement in an act is not accounted for by the fact that one has a second-order desire. One could have a second-order desire for something and still not act freely: for example, if that second-order desire is not a second-order volition. The only thing that makes a difference, then, is that her second-order desire to read Hegel is a second-order *volition*, while her second-order desire to go to the movies is not. Her second-order volition to read Hegel might be formulated as follows: "I want to have the desire to read Hegel, and to have that desire be my will." Strangely,

[47] Ibid 83.
[48] Ibid 86.

though, this just looks like a regular second-order desire ("I want to have the desire to read Hegel"), plus a first-order desire in disguise ("I want to actually read Hegel."). Thus, it seems that having a second-order volition simply means that you have a second-order desire to X, plus you *really* want to X. Frankfurt's hierarchy provides no way of explaining what it is to *really* want to X, so I do not think it helps. In fact, Frankfurt seems to agree: he appeals to a mechanism outside his hierarchical account: something called "decisive commitment."[49] With this he suggests that an agent must *really* commit to a course of action in order to act freely. Saying that an agent decisively commits to the particular hierarchy of desires which leads to action, unfortunately, is no different than saying that an agent is involved in the causal process leading from desire to action. If Frankfurt needs to appeal to such a thing, then he has not managed to give an account of the self in terms of beliefs and desires, so his hierarchical model cannot serve our purpose.

Gary Watson agrees that Frankfurt's model fails[50] and offers his own account as a replacement. I will call it the valuation account of the self. This account centers, unsurprisingly, on a set of mental phenomena that Watson calls "values". Watson does not define a value by itself, but he does define a "valuation system" as "…that set of considerations which, when combined with his factual beliefs (and probability estimates), yields judgments of the form: the thing for me to do in these circumstances, all things considered, is *a*." (Watson means this to contrast with one's motivational system, which is simply the set of things which move the agent to act). Soon after this, he says that "…

[49] Ibid 86-7.
[50] Gary Watson, "Free Agency," in *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001), 102-4.

[to] assign values to alternative states of affairs [is]…to rank them in terms of worth."[51] He also says that to value a state of affairs is to think it good.[52] That is not identical to wanting the state of affairs to obtain, Watson says, although "…to think something good is at the same time to desire it (or its promotion)."[53] Even though Watson only sidles up to the issue of defining a value, never quite making direct eye contact, his several proximal definitions work well enough. It appears that a value is supposed to be a judgment about whether a state of affairs is good and how good it is. Watson means 'good' to stand for 'good, all things considered.'

Values have structures similar to beliefs of the form: "*E*-ing is good and worthy," or perhaps, "*E*-ing is better and worthier than *A*-ing." Those are the kinds of judgments which could combine with ones desires to form the judgment, "The thing for me to do, considering these circumstances, is *E*." For some, perhaps it would not even require input from one's desires; although Watson discusses some values that are caused by one's desires,[54] some people's valuation systems may not contain any mention of their desires.

According to Watson, an agent acts freely (autonomously) when his "valuational system and motivational system…coincide. [When they do,]…what determines the agent's all-considered judgments also determines his actions."[55] In other words, an agent acts freely when her actions are caused by her values (as well as her desires). In terms of the standard story of action, then, the agent is involved when her values comprise at least some of the beliefs that, together with her desires, cause an intention, which in turn

[51] Ibid 100.
[52] Ibid 95.
[53] Ibid.
[54] Ibid 98.
[55] Ibid 100.

causes an action. The agent is represented in a causal chain by her beliefs about what is good.

In some cases Watson's position makes sense. Consider once again the case of the unwilling addict. She values her health, and her life, and various things with which taking drugs interferes. Sadly, she takes drugs anyway. Thus she acts contrary to her values, so her values do not make up part of the causal process leading to action, so she des not act freely when she takes the drugs. Watson's position gives us the answer we intuitively expect. Consider also the case of Frankfurt's wanton. A person who simply does what he desires without any reflection cannot properly be said to have values at all (or, at least, to have values that are ever causally efficacious). He never stops to think about the moral law, or even the best action, all things considered; he doesn't consider much of anything, let alone all things. Once again, Watson's position provides the result we expect: the wanton does not act freely, if the wanton can even be described as an agent. Of course, Watson's theory also provides the correct answer in a normal case, such as Susan's: assuming that Susan valued studying over relaxation, or something similar, we can say that her values played a role in causing her action, so she acted freely. If, for some reason, she actually valued relaxation over studying and read Hegel anyway, we would intuitively say that she did not act freely, just as Watson's position claims. Importantly, Watson can handle those cases without falling into regress or appealing to some mysterious "commitment" to action, so in that sense his account works better than Frankfurt's.

Watson's account might work for many commonplace actions, where the agent has an established value concerning the choice at hand and acts on it. A major problem

appears, however, when we consider a case where the agent doesn't have a standing value. How does the agent form a new value, a new belief about her best action? In Susan's case, for instance, how does she determine that she values studying and not watching the movie? She certainly can't just consult her ordinary beliefs and desires: those don't tell her how to rank her desires. She can't consult her values: by stipulation, she has no value that covers this situation.[56] She must appeal to something else, some other criterion, to determine her new value. That criterion, whatever it is, seems like a better candidate for counting as the agent. In actions that result in value formation, then, values alone cannot account for agential involvement. I imagine these value-forming actions serve the same purpose as Kane's SFAs, so the values formed in such a way count as agential involvement by proxy. Values might work as an intermediary for this unknown agent, then, but they cannot be the whole picture.

Both Watson and Frankfurt face another problem: their accounts don't make the self permanent enough. We may be willing to accept that people's selves change to some extent, but overall we expect them to remain more or less the same. When we say, "George is generous," we typically do not mean, "George is generous right now, but who knows what he was like before or will be like tomorrow." We typically mean, "George was, is, and probably will be generous." At the very least, we expect certain parts of ourselves to stay the same from one moment to the next. That is the part of ourselves that gives us our identity, the part that we praise or blame when we make moral judgments,[57]

[56] Ekstrom's account, described below, offers a way out of this problem. I'll discuss that when I get to her account.

[57] When we don't think that that is the part of a person fully responsible for an action (when she acts under compulsion, or from a fleeting but overwhelming desire (such as the case in "crimes of passion"), we at least mitigate our judgments of that person.

the part that we label our "true selves". As mentioned above, the true self needs to be causally involved in an action in order for a person to act autonomously.

Frankfurt mentions nothing about a need for second-order volitions to be permanent or semi-permanent features of a person's mind. It could be the case that Susan had nothing even approaching a second-order desire to read Hegel for her entire life. She may even have been an advocate for burning all of Hegel's books, and for daily attendance at the movie theater. If, for any reason, at the time of her decision, she suddenly had the second-order volition to read Hegel (assuming she "decisively committed" to it), a proponent of Frankfurt's theory would have to admit that she acted autonomously. From a common-sense perspective, however, such an act is not evidence of Susan's happily autonomous state, but of some kind of dangerous interference with the expression of her personality.

Watson's position is even worse on this count: Watson writes about values changing in response to environmental influences, and then changing back after that influence is removed.[58] If our "true selves" change when we are hungry or bored or when we find a dime in a phone booth[59], we will have to radically change how we see and treat ourselves. It would seem wrong to hold a person responsible for actions they took when they had a different true self, for instance. If we praise and blame the true self, and the true self changes, why should we praise or blame the new true self for the old one's behavior?

---

[58] Ibid 98-9.

[59] In *Lack of Character* (New York: Cambridge University Press, 2002) John Doris mentions a psychology experiment in which experimenters did or did not leave a dime in a phone booth on a public street. After a subject made a phone call (and did or did not find a dime), a confederate "accidentally" dropped her papers on the sidewalk in front of the subject. Of the 16 that found a dime, 14 helped the confederate gather her papers. Of the 25 that did not find a dime, one helped.

Maybe some philosophers would rather take that path. They would simply bite the bullet and say that we are confused about the self and about responsibility and need to revise our behaviors accordingly. I don't want to make that move, though. Remember the big assumption that I made in the introduction, that moral responsibility is possible. Even if someone proved that most people aren't morally responsible for most of their actions, my project wouldn't change. I'm interested in a conception of the self and of action that allows for moral responsibility. Frankfurt and Watson don't offer accounts that make the self permanent enough for responsibility, to their accounts don't serve my purpose.

Laura Ekstrom offers a view that can help deal with the permanency problem. She calls her theory, and so will I, a coherence theory of autonomy.[60] Ekstrom holds that the agent's self is composed of a certain subset of the agent's beliefs and desires. She calls the relevant desires "preferences"[61] and the beliefs "acceptances".[62] Preferences are second-order volitions "…formed in the search for what is good."[63] Acceptances are essentially values: they "…aid [the agent] in the activity of preference formation by indicating to him what sorts of actions and states of affairs are instrumentally and intrinsically good."[64] A person's self, then, consists of that person's acceptances and preferences, along with the power to "refashion"[65] that set.

The above describes Ekstrom's view of the self, but it describes the self in the broad sense, not the narrow sense we are after. Her broad definition of the self has some strange implications (nobody can have a self containing desires which they know to be

[60] Laura Waddell Ekstrom, "A Coherence Theory of Autonomy," *Philosophy and Phenomenological Research* Vol. LIII, No. 3, September 1993, 599.
[61] Ibid 603.
[62] Ibid 606.
[63] Ibid 603.
[64] Ibid 606.
[65] Ibid.

bad, or even morally neutral), but I mention it only as a step towards the "true self". According to Ekstrom, the "true self" is "…a subset of these acceptances and preferences, namely, those that *cohere* together."[66] Ekstrom calls those preferences and acceptances *authorized*, and offers three reasons to accept her proposal: First, a coherent set tends to be long-lasting, as we would expect a person's self to be. The true self is not fickle, as Watson's valuation system is. Second, authorized preferences and acceptances offer each other mutual support, mutual justification. One's true self should consist of those preferences and acceptances most fervently held, those that hold up to the most challenge and scrutiny. Since the members of a coherent set provide each other with justification, they are much more likely to hold up to such a challenge. In addition, each authorized preference and acceptance, since it is consistent with the rest of the members of the set, better represents the rest of one's self than a one-off preference inconsistent with many of one's other preferences and acceptances. Third, one's authorized preferences and acceptances are those upon which one is comfortable acting. Since they cohere with a number of one's other acceptances and preferences, one will not feel conflicted when acting upon them.

According to Ekstrom's theory one acts autonomously when "…one acts on a first-level desire because one has a personally authorized preference for that to be one's effective desire."[67] She claims further that the action must be *caused* by that preference, in order to rule out an action for which one has a personally authorized preference being caused by an external force. Thus, her theory involves the agent in the standard story of action by requiring the involvement of a personally authorized preference. Perhaps the

[66] Ibid 608.
[67] Ibid 614.

story goes something like this: There is something that the agent wants, and there is an action that he believes conducive to its attainment. His desire for the end, his belief in the action as a means, and his personally authorized preference for that desire, justify taking the action, and they jointly cause an intention to take it, which in turn causes the corresponding movements of the agent's body. The desire and belief may or may not be a part of the agent's self or true self, so they do not represent the agent. Only the personally authorized preference does so. Such an account leads to the proper conclusions in our basic test cases: in the unwilling addict doesn't act autonomously, and Susan, in choosing to read Hegel, does act autonomously.

Ekstrom isn't clear about whether she means 'good' to stand for the morally good, the prudentially good, or the "good, all things considered," but she's in trouble regardless. Her account cannot fully overcome Frankfurt's and Watson's problems. She can overcome the "first order" of the problems: individual preferences are formed on the basis of coherency with the person's existing set of preferences and acceptances. That ends the regress and provides a criterion for preference formation. It's only a partial fix, though: much like the supporters of the third intuition concerning moral responsibility, she faces a problem with origin of the self. How does the agent form her first preference? She doesn't have any set of existing preferences or acceptances on which to rely. Just like Watson's agent trying to form values, or a "little agent" facing her first significant decision, Ekstrom's agent has no criterion for choice. We have to assume that there is such a criterion in order to get her account off the ground, but that makes most of Ekstrom's system superfluous. The agent's preferences and acceptances should just

cohere with that original criterion of choice, whatever it is. They'll cohere with one another, too, but that's beside the point.

After examining those three attempts to present the self in terms of beliefs and desires it seems clear that those mental states aren't the right tools for the job. We need some other mental phenomenon to capture, or at least represent, the self in the narrow sense. We need another kind of thing involved in the process of action in order for that action to count as autonomous.

We can make a list of requirements for anything that could constitute or represent the "narrow" self by rephrasing the deficiencies of the above theories. It seems there are four requirements. First, the thing must remain relatively unchanged through time. Second, the thing must provide a criterion of choice between opposing desires or beliefs. Third, the thing must be involved (assuming the existence of the thing) in all or most actions we intuitively consider autonomous. Fourth, the thing must not be involved in all or most actions we intuitively do not consider autonomous. In the next section I outline an account that meets all of those requirements, centered on a mental object that I call the preference set.

III. The Preference Set

Since I've rejected beliefs and desires as adequate representatives of the true self, I'll explain the preference set by contrasting it with those things. First, unlike beliefs and desires, the preference set doesn't change once it's fully formed. That means it's probably not a mental state, since mental states correspond to temporary neuronal firings in the brain. I'll instead call the preference set a 'mental structure' to emphasize its

permanence. I imagine that mental structures probably correspond, in beings like us, to some configuration of neurons and glial cells in our brains. In some cases, they might consist of single cells connected in the proper way to other parts of the brain. I don't want to define mental structures in terms of those cells, though, because it seems likely to me that a thing made of something other than neurons could have mental structures.

Instead, I'll define mental structures functionally: a mental structure is a mental object that performs a particular mental function. For instance, the thalamus is a kind of mental structure: it receives information from the senses, reorganizes it, and sends it to the parts of the brain that process it. The simple cells in the primary visual cortex are another kind of mental structure: they respond to lines within the visual field, and respond to changes in those lines, and so allow us to detect motion. The preference set is also a kind of mental structure: it takes as input beliefs and desires and produces intentions. It consists of simpler mental structures that I will call, unsurprisingly, 'preferences.' Like the thalamus and the simple cells in the primary visual cortex, the preference set probably forms as part of the normal expression of a person's genes. People probably aren't born with a fully-formed preference set: my guess is that it mostly finishes forming around the age of reason (usually about 7 years old)[68]. It may take longer, though: many scientists think that people's brains aren't mature until their late twenties[69].

Beliefs and desires are relatively simple things: they usually concern single states of affairs. For example, one might believe that tasting chocolate is a pleasurable state of

[68] Adele M. Brodkin, "The Age of Reason," *Scholastic Parents* (July 1, 2006), http://www2.scholastic.com/browse/article.jsp?id=7241, para. 3.
[69] Deborah Bradley Ruder, "The Teen Brain," *Harvard Magazine* (September-October 2008), http://harvardmagazine.com/2008/09/the-teen-brain.html, para. 5.

affairs, and so desire have a desire for a state of affairs in which one tastes chocolate. Preference sets, on the other hand, are very complex things, consisting of a particular hierarchy of preferences. In fact, it might be misleading to refer to individual preferences, because preferences have no real importance individually. They aren't the preferences Ekstrom discusses. They are closer to the preferences used in economics, and are significant only in their relations to one another. Each preference corresponds to a particular feeling or sensation, or a combination of several. For example, the feeling of asserting oneself, of safety, and of doing what one ought to do (according to the moral law)[70] might comprise three preferences.

Each preference is ranked with respect to each other preference in one's preference set, allowing for no ties. The preference set is therefore a single-peaked system: it provides a clear hierarchy of states of affairs, unambiguously ranking one randomly-chosen set of feelings over any other randomly picked one. For example, a preference set with just the three variables given above might look like the following:

1. One acts assertively, safely, and in accordance with the moral law.
2. One acts assertively, recklessly, and in accordance with the moral law.
3. One acts passively, safely, and contrary to the moral law.
4. One acts passively, recklessly, and contrary to the moral law.
5. One acts assertively, safely, and contrary to the moral law.
6. One acts passively, safely, and in accordance with the moral law.
7. One acts passively, recklessly, and in accordance with the moral law.
8. One acts assertively, recklessly, and contrary to the moral law.

Note that no particular feeling is "more important" than any other feeling: the person doesn't always prefer states of affairs in which she acts assertively to states of

[70] These are obviously very specific feelings, and some people might not have such fine distinctions in their preferences. For instance, one might not feel any difference between doing the morally right thing and the socially acceptable thing. In that case, I suppose decisions that involve either of those things would activate the same part of the agent's preference set.

affairs in which she doesn't. That kind of preference set might exist; for example, a moral saint would prefer all states of affairs in which she acts in accordance with the moral law to all states of affairs in which she doesn't. However, as in the three-preference example above, it probably doesn't work that way. A preference set allows for any order of rankings.

That allows for enormous variation between sets with the same variables, increasing exponentially (2^n, where n is the number of variables) with each variable. In a set including 20 variables, then, there are 1,048,576 possible hierarchies. In one with 50, there are over a trillion. Since humans likely have a huge number of variables in our preference sets (a huge number of gradations in feelings and combinations of feelings), that makes it extremely unlikely (though not impossible) that we have any "repeat" people. Besides, even in that ridiculously improbable case, the preference sets would be numerically distinct, so we would have no problem distinguishing between two different people.

As mentioned above, the preference set takes as input beliefs and desires and creates intentions. I imagine this works in a way similar to the following. The person has a certain set of beliefs about which states of affairs will cause certain feelings or sensations, and about which of those states of affairs the person can act to bring about. Through some mental process that we call 'deciding,' the information contained in those beliefs is fed to the preference set, which determines which possible state of affairs provides the highest-ranked feeling or set of feelings[71]. It then causes the person to form an intention to act in order to bring about that state of affairs.

[71] This process probably also accounts for the probability of bringing about the various states of affairs and their corresponding feelings.

To rephrase it in similar terms to the standard story of action:

> There is something an agent *prefers*, and an action he believes conducive to its attainment. His *preference* for the end, and his belief in the action as a means, justify taking the action, and his preference set causes an intention to take it, which in turn causes the corresponding motions of the agent's body. Provided that these causal processes take their normal course, the agent's movements consummate an action, and his motivating *preference* and belief constitute his reasons for acting[72].

Some readers might be confused, wondering (a) what happened to desires, and (b) how preferences are significantly different from desires anyway. At first glance, the two phenomena look fairly similar: they both make us want, in some sense, for a particular state of affairs to obtain. That seeming similarity is misleading, though. Individual preferences have no strength or force. They are only meaningful in reference to other preferences. If I say that I had a preference for eating cheeseburgers, I really mean that my preference for (the sensation I associate with) eating cheeseburgers is ranked higher than my preferences for (the sensation I associate with) eating many other foods. I don't mean that I desire a cheeseburger; I might have no urge of any particular strength to get one and eat it. I may not even enjoy the feeling of eating a cheeseburger. Maybe I don't enjoy eating at all, and prefer cheeseburgers as the least of many evils.

Desires serve one of two functions in my model. First, they might cause a person to have certain beliefs about himself, which would affect which parts of his preference set got consulted in making a decision. For instance, I might think that I desire a cheeseburger, and so expect a different sensation (i.e. the satisfaction of that desire) when

[72] Any action proceeding from such an intention, including long processes or components of a larger action, should also count as autonomous. For example, I might decide before a soccer game that I am going to play my best and focus on team play. Even if each action I perform in the course of the game is not the result of its own individual autonomously formed intention, the individual intentions causing the actions are themselves caused by an autonomously-formed intention. The important thing is that the actions have an autonomously-formed intention somewhere in its causal history.

I eat a cheeseburger than if I didn't desire one. I'll also probably rank my current, cheeseburgerless state lower than I would if I didn't have a desire for a cheeseburger. In other words, my desire causes me to associate a different part of my preference set with the current state of affairs (and different parts with potential future states of affairs), and so affects the intention created by my preference set Desires seem to function in that way most of the time.

Second, in some rare cases desires lead to intention formation themselves, bypassing the preference set entirely. I'm not sure exactly how this works. Maybe they somehow suppress the causal mechanism that involves the preference set. That seems to fit the case of the unwilling addict fairly well. If we think that consulting one's preference set constitutes self-expression or self-control, having a desire that bypasses the preference set would make one feel coerced and helpless, just like the addict.

IV. The Preference Set as the Self

In this section I argue that the preference set is just that: the self. Why accept such a claim? Simply put, the preference set satisfies the four requirements stated for anything attempting to represent the self. First, it remains unchanged through time. In fact, barring direct physical manipulation of the brain, the preference set cannot change at all. Just like the thalamus or the complex cells in the primary visual cortex, the preference set corresponds to some clump of neurons in the brain[73]. That clump formed in accordance with that person's DNA and normal development, so when Abraham Lincoln gave the

[73] Again, a different kind of creature might have a different kind of structure that serves the same purpose, in which case I would still consider it a preference set. In things like us, though, the preference set would have to be made of neurons.

Gettysburg Address he had the same preference set that he had when Booth shot him.[74] A person's beliefs and desires can change, of course. For instance, it seems possible that an extremely assertive person could end up with a sort of "learned helplessness". She might jump at the chance of asserting herself if she only had the chance. Unfortunately, if she believed that no situation she ever faced presented an opportunity for asserting herself, she would never act assertively. That is not to say that she doesn't act on her preferences, only that the causally efficacious preferences never include those states of affairs in which she acted assertively. She could face plenty of opportunities to assert herself, but be tragically deluded about the consequences of her possible actions. The same might occur to a moral saint: he might always act in accordance with the perceived moral law, but if he were mistaken about that moral law, he might act against the moral law all the time.

My account doesn't seem to leave any room for self-improvement, for building character. It seems that self-improvement would require changes to the preference set, which my theory says is impossible. However, self-improvement can also be represented as a person's desires shifting to better match one's preference set. For example, one of my friends is a great philanthropist, although he was not always one. When he first started volunteering he hated every minute of it. It was a struggle to perform what he saw as a necessary service. Over time, however, he grew used to volunteering, and now he finds it fun. That does not mean that his preference set has changed. In fact, it seems clear that he always preferred to engage in philanthropy, since he did it even when he hated it. The change is better represented as occurring at the level of belief and desire: the

[74] His preference set was functionally the same, even if it consisted of numerically different neurons.

combination of feelings he got when he acted charitably changed, as did his expectations about how philanthropy would make him feel.

One might object that many people seem to, over the course of their lives, change the order of their preferences. Remember the simplified three-component preference set I discussed earlier. It seems quite possible for a person with that preference set to change her mind, such that, for instance, she always acted morally (or at least always tried). Maybe she read Kant and he convinced her that the moral law was objective and real and more important than any prudential considerations. Surely, in that case, her preference set has changed.

The problem this objection poses lies not with my position, but with my example. I used a three-component preference set because I would have found it tedious to write out a 200-component preference set. I think the change in our hypothetical agent's actions is best explained by appeal to another part of her preference set. After all, we need to explain why she changed her behavior. Her reading Kant serves as a partial explanation, but not everyone who reads Kant, not even everyone who finds Kant convincing, tries to act morally all the time. Why did our agent react that way? Well, it seems that she has a preference for acting in accordance with theories that she thinks are true.[75] Our simplified version of her preference set didn't take that part of her set into account. It consequently looked like she changed her preferences, when really the change in her beliefs caused the same kinds of situations to activate a different part of her preference set than those kinds of situations previously activated.

[75] Not everyone has such a preference. I know plenty of people who think that eating meat is morally wrong and that they have a moral obligation not to eat it, yet keep eating meat.

What about more radical changes to personality, such as religious conversions and the like? I actually find it kind of surprising that people think of religious conversions as real changes in a person's character. Why not think instead that the person has acquired a new belief, one that pertains to nearly every state of affairs and alters the feelings that the person feels in those states of affairs? For instance, if a person suddenly acquires a strong belief in God, she will act in ways that she thinks are in accordance with God's will. That's not because her personality has changed, but because she thinks her actions have different consequences. She might think that being kind to neighbors and respectful to her parents grants her eternal life, or gets her God's love, or just serves as an expression of her faith. Those acts will consequently trigger different feelings in her, so they will involve different parts of her preference set.

The preference set also meets the second requirement for any mental phenomenon representing the self: it provides a criterion of choice between various desires and beliefs. It provides a non-arbitrary criterion for choice between any two states of affairs that correspond to different preferences.[76] Desires and beliefs might provide such a criterion between two *particular* states of affairs, but they cannot generalize like the preference set: they cannot provide a reason for choosing between competing beliefs and desires (at least, not without a vicious regress). The model meets the third requirement as well: it fits our intuitions concerning which actions are autonomous and which are not. As shown above, the model gives a perfectly good account of how the agent is involved in a regular

[76] To be more precise, it provides a non-arbitrary criterion for choice between any two states of affairs that the agent *believes* correspond to different preferences. It is also possible, although extremely unlikely (given 50 variables, less than one in a trillion situations), for two states of affairs to correspond to the same preference, which is the only case in which the preference set cannot provide a criterion of choice. In that situation a person might just say, "I really, genuinely don't care either way." I imagine we have some sort of pseudorandom mechanism for determining our course of action in such cases, so that we do not end up like Buridan's Ass.

autonomous action: the preference set is part of the causal process leading to the action. Unlike Watson's and Ekstrom's theories, it does not exclude cases where an agent intentionally does wrong. The agent might consider the feeling associated with following the moral law relatively unimportant, and so act in accordance with her preference set.

What about the fourth requirement, that the thing not be involved in actions that we intuitively consider non-autonomous? As mentioned above, in the cases of the unwilling addict and the wanton the actions come out non-autonomous, as we would expect. Those who have experienced extreme trauma likely act non-autonomously as well, because their actions seem to be caused by unreflective impulses or desires (that have no autonomously-formed intention in their causal histories). Direct manipulation of desires or the causal process leading to action also violates the requirements for autonomy, since it excludes the involvement of the preference set and so the self. The case of manipulation of beliefs takes a bit more work to fit into the theory. According to the theory, autonomy doesn't depend at all on what a person believes, so long as those beliefs interact properly with the preference set. This means that people who are totally deluded about the world nevertheless might act autonomously. Members of suicide cults, schizophrenics, even external world skeptics all act autonomously on this view. Some might see that result as a problem, since we tend to treat delusional people quite differently than regular people with respect to their actions. However, the problem dissolves when one considers the difference between autonomy and responsibility.

Responsible actions generally have an epistemic requirement[77], so we can say that delusional people act autonomously but not responsibly.

Only manipulation of the preference set itself presents any real problem for my theory. As mentioned, such manipulation cannot be accomplished by affecting a person's desires or beliefs, since the preference does not change along with them. The preference set is consequently impervious to most kinds of manipulation. However, since each preference set corresponds to a particular configuration of brain cells, one way to affect a person's preference set exists: poke him in the brain! If the pertinent part of a person's brain is damaged then their preference set changes, according to the theory. If such a thing occurs, can the person still act autonomously? My answer is no, but for a somewhat strange reason. Remember that the preference set just is the self, and the self just is the preference set. That means that any change to the preference set is a change to the self. Since the preference set after the damage is a different preference set than the preference set before the damage, the person after the damage is a different person than the one before the damage. For example, if Mary takes damage to her preference set and it

---

[77] For instance, George Sher defends the following condition, (quoted in Neal Tognazzini, review of *Who Knew? Responisibility Without Awareness*, by George Sher, *Notre Dame Philosophical Reviews*, January 3, 2010) :

> When someone performs an act in a way that satisfies the voluntariness condition, and when he also satisfies any other conditions for responsibility that are independent of the epistemic condition, he is responsible for his act's morally or prudentially relevant feature if, but only if, he either
>
> (1) is consciously aware that the act has that feature (i.e., is wrong or foolish or right or prudent) when he performs it; or else
>
> (2) is unaware that the act is wrong or foolish despite having evidence for its wrongness or foolishness his failure to recognize which
>
> (a) falls below some applicable standard, and
>
> (b) is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits; or else
>
> (3) is unaware that the act is right or prudent despite having made enough cognitive contact with the evidence for its rightness or prudence to enable him to perform the act on that basis.

changes as a result, Mary herself *is destroyed*. Her body continues to function, of course, but a different self, a different *agent*, inhabits it. The new person, Mary 2, may still act autonomously, of course, but Mary 1 cannot, because she doesn't exist anymore. That is at first glance a very strange result, but some precedent for it exists. There are many cases where people have taken brain damage and become "completely different people". The same is true of many patients with degenerative brain diseases such as Alzheimer's and dementia (though not all of them, of course). Families and friends often struggle to maintain their relationships with such people, since they undergo a serious personality change. It might be counterintuitive to say that the person is totally obliterated, since much of them remains. For example, the person might retain all or most of her memory after a brain injury. However, this intuition seems to be the result of the ambiguity of "self". While true that the person's broadly-understood self might remain, it is clear that her narrowly-understood self, her "true self", no longer exists.

V. Conclusion

I have tried to remove some of the vagueness that obscures the intuitions in play in the free will debate. By assuming that moral responsibility is logically possible, I was able to examine and evaluate three basic intuitions about moral responsibility. I preserved as much of the original intuition as I could without allowing for any impossible conditions. That process led to a set of conditions that appear compatible with determinism and incompatible with indeterminism. If I have offered sound arguments, then, it seems that the would-be defenders of moral responsibility should adopt compatibilist, rather than incompatibilist, assumptions. Even if I haven't, I hope that at

least I've helped to clarify just what the argument between libertarians and compatibilists is about. That will help both sides focus on the points under contention, rather than waste time on arguments that rely on different intuitions, and so have no force for anyone but those who already accept them.

I also tried to offer an account of the self, and of the self's role in action, that is compatible with moral responsibility. Opponents in the free will debate often seem to have different ideas about the self, as well, which can also get in the way of productive argument. My preference set theory seems at least a plausible account of the nature of the self, one that all parties in the debate could use to avoid that distraction. It captures our intuitions about the self and allows for an account of autonomous action that properly includes the involvement of the self, something opposing theories cannot do. It also seems useful in solving problems related to moral responsibility, especially for compatibilists. At the very least, it is a step in the right direction for the debate about autonomy, as well as the debate about moral responsibility: it provides a clear account of the self, something both debates sorely lack at present.

Bibliography

Brodkin, Adele M. "The Age of Reason." *Scholastic Parents* (July 1, 2006), http://www2.scholastic.com/browse/article.jsp?id=7241.

Chisolm, Roderick. "Human Freedom and the Self." In *Agency and Responsibility*, ed. Laura Waddell Ekstrom, 126-137. United States of America: Westview Press, 2001.

Doris, John. *Lack of Character.* New York: Cambridge University Press, 2002.

Ekstrom, Laura Waddell. "A Coherence Theory of Autonomy." *Philosophy and Phenomenological Research* Vol. LIII, No. 3, September 1993: 599-616.

Ekstrom, Laura Waddell. "Indeterminist Free Action." In *Agency and Responsibility*, ed. Laura Waddell Ekstrom (United States of America: Westview Press, 2001): 138-157.

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (1969): 829-839.

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." In *Agency and Responsibility*, ed. Laura Waddell Ekstrom, 77-91. United States of America: Westview Press, 2001. Originally published in *Journal of Philosophy* LXVIII, 1 (January 1971): 5-20.

Kane, Robert. "Responsibility, Luck and Chance." In *Agency and Responsibility*, ed. Laura Waddell Ekstrom, 158-180. United States of America: Westview Press, 2001. Originally published in *Journal of Philosophy* XCVI, 5 (May 1995): 217-240.

Mele, Alfred R. *Free Will and Luck*. New York: Oxford University Press, 2006.

Ruder, Deborah Bradley. "The Teen Brain." *Harvard Magazine* (September-October 2008), http://harvardmagazine.com/2008/09/the-teen-brain.html.

Strawson, Galen. "Luck Swallows Everything." *Times Literary Supplement*, June 26, 1998.

Strawson, Peter. "Freedom and Resentment." In *Agency and Responsibility*, ed. Laura Waddell Ekstrom, 183-204. United States of America: Westview Press, 2001. Originally published in *Proceedings of the British Academy*, Vol. 48 (1962).

Tognazzini, Neal. Review of *Who Knew? Responisibility Without Awareness*, by George Sher, *Notre Dame Philosophical Reviews*. January 3, 2010.

Velleman, J. David. "What Happens When Someone Acts?" *Mind* 101, 403 (1992): 461-

481.

Watson, Gary. “Free Agency.” In *Agency and Responsibility*, ed. Laura Waddell Ekstrom, 92-106. United States of America: Westview Press, 2001. Originally published in *Journal of Philosophy* LXXII, 8 (April 1975): 205-220.