Stony Brook University

An Unconditioned Will: The Role of Temporality in Freedom and Agency

A Dissertation Presented

by

Roman Altshuler

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Philosophy

Stony Brook University

May 2010

Stony Brook University

The Graduate School

Roman Altshuler

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Dr. Don Ihde – Dissertation Advisor
Distinguished Professor of Philosophy

Dr. Donn Welton – Chairperson of Defense
Professor of Philosophy

Dr. Jeffrey Edwards
Associate Professor of Philosophy

Dr. Neil Levy
Director of Research
Oxford Centre for Neuroethics

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

An Unconditioned Will: The Role of Temporality in Freedom and Agency

by

Roman Altshuler

Doctor of Philosophy

in

Philosophy

Stony Brook University

2010

Eliminativists about free will and moral responsibility argue that no action can be free and responsible because in order to be actions, our movements must be caused by features of our character or will. However, either the will is constituted by states that are themselves produced by events outside our control, or it is constituted by our own choices, which must themselves stem from our will in order to be up to us. Thus, any attempt to account for freedom and responsibility seems to either run into an infinite regress or leave the ultimate causes of our actions up to something outside our agency. Compatibilists attempt to respond to this challenge by arguing that we need not have control over our will in order to be free, but only to have control of our actions on the basis of our will. Libertarians, on the other hand, argue that we can be free so long as our choices are caused indeterministically and chosen for reasons.

I argue that both approaches ultimately leave the constitution of the will up to non-agential factors because the dominant accounts view all choices—including those that constitute the will—as essentially events caused by other events, leaving no function for agents to perform. In response, I argue that we can avoid eliminativism if we take the will to be irreducible to events such as choices and also our *own*. Through an examination of recent non-volitionist approaches that allow for responsibility for non-deliberative action, I argue that such accounts presuppose a Heideggerian view of agency on which all action and deliberation occur on the basis of an underlying projection of possibilities into which we are thrown. Heidegger's account of temporality in turn allows us to own ourselves in the present by retrieving our past as always already chosen in light of our self-projection into the future. Agents are thus self-constituting beings capable of owning themselves and independent of causation by prior events. Freedom and responsibility are therefore irreducible features of agency.

Table of Contents

1. Introduction ……1
A. Motivations and Threats ……2
B. Shallow and Deep Temporality …….....5
C. Summary of the Dissertation …….....8

2. Free Will …… 10
A. Eliminativism …… 10
B. Compatibilism …….....24
I. Davidson…….....27
II. Frankfurt ……..38
C. Libertarianism ……... 49

3. Experience, Deliberation, and the Construction of the Will ……...62
A. Searle ……62
B. Korsgaard ……..... 70
C. The Will ……80
D. A First Shot at a Solution ……..87
E. Time and Ownership …… 96

4. Responsibility ……..... 100
A. The Impossibility of Moral Responsibility ……... 100
B. Responsibility and Consciously Thematized Decision……... 111
C. Defending and Redefining Attributionism …… 133

5. Control and History …… 149
A. Control ……..... 149
B. History and Ownership …… 186

6. Temporality and Ownership ……... 217
A. Discovering Ownership …… 217
B. Constitutivism: Setting the Stage …….. 222
C. Conscience and Resoluteness ……234
D. Death and Being-a-Whole ……246
E. Anticipation and Temporality ……258
F. The Situation and Self-Ownership ……269

7. Conclusion ……...279

References ……...284

1

Introduction

> It is impossible for the will, which cannot of its very nature do otherwise than obey itself (for there is none who doth not will what he willeth, or who willeth what he doth not will), to be deprived of its freedom. The will can, indeed, be changed, but only to another will,- in such a way that it never loseth its freedom. Therefore it can no more be deprived of its freedom than it can be deprived of itself.
>
> — Bernard of Clairvaux

> Our will would not be a will if it were not in our power. And since it is in our power, we are free with respect to it.
>
> — Augustine

The Medievals believed that the will is free by definition. To be an agent—to have a will—just *is* to be free on their account. But this is not the view held by most contemporary philosophers. Instead, the tendency is to take up a theory of agency and then ask whether free will is a property that agents, so defined, can have. At least part of my argument will be to suggest that this is a mistake. If free will is separated from agency in this way, it loses coherence. I will focus primarily on this conceptual problem: how can we even make sense of the idea of a free will? Drawing on Nagel's argument for the incoherence of free will, I will argue that the problem is even more complex than usually assumed. The difficulty of understanding free will is not simply one of figuring out whether nondeterministic causation is possible, or of trying to match up our commitments with our actions in just the right way. Rather, I will argue, the difficulty involves trying to make sense of how the movements of some region of the world could be assigned

selfhood, agency, or freedom. In the rest of this introduction, I will briefly note the motivation for developing a theory of free will and the threat to it, introduce the idea of deep temporality that I will use throughout, and provide a brief summary of the following chapters.

A. Motivations and Threats

What are the motivations for defending free will? In explaining the motivations for a libertarian view of free will, Robert Kane outlines a number of things we might want that seem to require that we be the originators of our actions. Among these, he lists creativity, autonomy, desert, moral responsibility, being an appropriate object for reactive attitudes, dignity, individuality, life-hopes, and genuine love or friendship. (Kane 1996 80 ff.) I confess that for many of these, I lack the intuition that they depend on any kind of free will. For others, the sort of free will they require does not seem to demand anything as extreme as origination. For some, the "challenge" supposedly posed by the absence of free will, or at least of origination, seems to be simply the problem posed by the possibility of prediction. Someone might be outraged, for example, if they were told that their wonderfully inventive piano solo could have been predicted in detail even before their birth. But this is hardly a pressing problem: whether someone might dislike the idea that their actions are predictable has no bearing on free will unless it is at least *possible* to predict the future with absolute certainty given the current state of the universe. But there are reasons to doubt whether this is even a conceptual possibility (at least for any entity that exists in time rather than eternity), and it is difficult to imagine it as a real

technological possibility. In any case, concerns that prediction could frustrate some of our hopes or desires seems only tangentially connected with the issue of free will.

On the other hand, even the emphasis on our hopes or desires strikes me as misguided. Philosophy is not—or should not be—a discipline of developing theories to explain the world so as to match our desires. That we desire free will for one purpose or another is neither here nor there, especially since we need hardly worry that eliminating free will is likely to end things like genuine friendship or love—even were hard determinism and all its consequences fully accepted by the general population, friendship and love would no doubt go on, with or without an implicit change in the implications involving them. Something more than a desire seems needed to provide a motivation for free will. In this regard I want to emphasize moral responsibility. It is true that, historically and to this day, one motive for developing theories of free will is a retributivist one: those interested in defending various practices of punishment, for example, may have a desire to hold others responsible. In this sense, the search for free will once again involves only a desire to justify existing institutions, ones without which we could likely continue to live quite well, and perhaps even better.

But there is also a sense of moral responsibility that has a stronger appeal, one that goes beyond simple desire. I doubt we can get something important out of responsibility taken as a third-personal concept; but as a first-personal one, it is connected to something more significant: moral self-determination. While to go into the topic in detail would derail the purposes of an introduction, I see a theory of free will as providing a foundation for moral philosophy. To postulate moral laws as *unconditionally* binding, and to understand ourselves as able to respond to them, we need a theory of free will as

an *unconditioned* will. Of course even without freedom of any sort we can strive to follow moral laws. But without an unconditioned will, our success or failure will be only a matter of luck; it will not be attributable to us. And in this sense, the search for free will as necessary for moral responsibility as self-determination is not motivated simply by desire. If morality *commands* us not merely to act in certain ways, but to act in those ways on the basis of an unconditional acceptance of the moral motives as our sole motives, we will need a free will to be able to fulfill our obligation. Of course this view of morality is not especially popular in Anglophone ethics. It is simply the motivation for seeking a coherent account of free will that I find especially compelling: we need free will not in order to satisfy a desire to be moral agents, but in order to do our moral duty.

What is the threat to this sort of free will? In the course of the past few centuries, but going back in time at least to the Stoics, the major threat to free will has been taken to be causal determinism. Causal determinism is the thesis that the state of the universe at any time is necessitated by the state of the universe at any previous time, together with the laws of nature. Consequently, much ink has been spent on defending a space for indeterminism within nature. But other philosophers have been skeptical of the idea that causal determinism as such is the real threat to free will. Thomas Scanlon, for example, dubs the main threat to free will the Causal Thesis:

> This is the thesis that the events which are human actions, thoughts, and decisions are linked to antecedent events by causal laws as deterministic as those governing other goings-on in the universe. According to this thesis, given antecedent conditions and the laws of nature, the occurrence of an act of a specific kind follows, either with certainty or with a certain degree of probability, the indeterminacy being due to chance factors of the sort involved in other natural processes. (1988 152)

This thesis differs from determinism because it recognizes that indeterminism alone does not provide any obvious help to defenders of free will. If our actions are not causally necessitated but merely caused, this cannot help give us free will if the indeterminacy involved in free will is of the same kind as the indeterminacy involved in any other, non-agential events. It is something very much like the Causal Thesis that Nagel takes up in his attack on free will. Throughout, I will generally use the term "eliminativism" to describe this threat after introducing it in Chapter 2A. At times, however, I will simply use "determinism," with the caveat that causal determinism is not my only, or even primary concern. After all, if—as the Causal Thesis suggests—my actions are no different in terms of their production than any other event in the physical world, the problem seems to arise that my actions are not attributable to *me* in any genuine sense; they are not products of my agency, unless agency is merely the name of a more ordinary type of causal process. And if my actions do not spring from *me*, and whether or not they happen depends ultimately on events prior to any exercise of agency on my part, then my actions are, for all intents and purposes, determined by non-agential factors.

B. Shallow and Deep Temporality

I will describe shallow and deep temporality in terms of the temporal relation each account allows between two entities. I will illustrate with a few examples. Imagine that I am a sniper lying in ambush behind a rock, waiting for a caravan to go past below me so I can startle the horses and pacify the guards. If I shoot too soon, they will return fire and avoid the disaster I intend for them; if I wait too long, they will be able to escape.

I wait for the perfect moment, or at least the moment that seems perfect to me, and deliberately pull the trigger. Here I have already made either a somewhat complex choice, or a complex of two different choices: I have decided to shoot, and I have decided on the exact moment at which I will shoot. In any case, this one decision, or this complex of decisions, is made at an identifiable moment in time. This need not mean, of course, that we can always pinpoint the moment with certainty, or that the moment is literally momentary. The point, rather, is that it is at least in principle a choice that we can plot out as an event on a time-line, maybe placing it a few micro-seconds before the movement of my finger on the trigger and perhaps half a minute after the first riders of the caravan enter my visual field. This choice (or complex of choices), is thus an event, identifiable in time in relation to other events.

Suppose, now, that I am sitting in a bar. I have called my partner to tell her not to wait up, because I expected to be working late. I managed to finish work early, however, and decided to wind-down over a drink. A brunette approaches the bar and begins speaking to me in riddles with a tantalizingly deep voice. Within this increasingly clichéd and fully fictional scenario, suppose that I have a choice to make: I can finish my drink and go home to surprise my partner by returning in time for supper, or I can make off with my new acquaintance. If I am presented with such a choice, I cannot be presented with it ex-nihilo. It presents itself as a choice for me because there is some at least potential internal conflict, something in the way of my deciding effortlessly one way or the other. For example, I have decided to be faithful to my partner and consider myself bound by that decision (otherwise, what kind of decision would it be?). On the other hand, in order to have a conflict, there must be something tempting me in the other

direction—a certain weakness for mysterious women, for example. The conflict here is between two opposing attitudes or constituents of my will. It is only on the basis of a will that I can face—and make—choices at all. The constituents of my will, moreover, range across a number of choices or possible choices. They are therefore not simple events, but cover entire time-spans containing events.

If I did indeed at some point decide to be faithful to my partner, then that attitude itself results from a choice made at some point in time; that is, such a constituent of the will is *reducible to an event*. On the other hand, it is possible that I have always been susceptible to mysterious women, and that no origin point in time can be found for this susceptibility; that attitude, then, is something that is present within the time-frame of all my choices; in a sense, it precedes any choices I might make on its basis and cannot be mapped out relative to them on a timeline. When we say, for example, that lying has always come naturally to Jack, that Jill has never been able to commit herself, or that Snip has never been partial to candy corn, we imply a temporality of this sort. A classic example here is the notion of an instinct in Freud's sense. Freud speculated that the two competing instincts, Eros and the death instinct, are not simply psychic entities—they are already built into any living entity, no matter how simple. (Freud 1960) The instincts are will-constituting attitudes in my sense, since they serve as underlying influences on our choices. But since to be human, to be alive at all, is already to have these instincts, it is not through occurring (as events in time), but through existing that they serve their function as constituents of the will. They are the sort of attitudes that do not belong on a timeline along with the events of our choices. In other words—and nothing is meant to ride on this example—our wills make possible and structure the events of choosing, but

the constituents of our wills may or may not themselves be reducible to prior events, whether of choice or not.

I will call an account temporally shallow when all the elements among which it postulates relations are either events or reducible to events. For example, if I choose to be a clown and then, on the basis of this attitude, choose to make a joke at every possible occasion, then the relation here is a temporally shallow one. Similarly, if my actions are caused by my will, which itself is the causal result of other events stretching back into the distant past, this account is temporally shallow. It is shallow because every element in the account can be plotted on the same timeline as occurring earlier or later than another element. On the other hand, a temporally deep account will include at least one element—perhaps the will, or some of its constituent attitudes—that is neither an event nor reducible to an event. To anticipate: I will argue that free will requires a temporally deep account on which our will is not reducible to any events and is also *our own*.

C. Summary of the Dissertation

I will split up free will and moral responsibility into two discussions, each addressed in two chapters. Chapters 2 and 3 focus on free will. In chapter 2, I work out Nagel's eliminativism and go on to suggest that neither compatibilist nor libertarian strategies are likely to work against it. The aim here will not be to examine cutting edge work on free will, and my argument is not meant to be a knock-down attack on all attempts to resolve the free will problem. Both contemporary compatibilist and libertarian accounts are frequently much more subtle and involved than my comments here will

suggest. My aim is only to make a plausible case for the view that eliminativism cannot be overcome along the usual lines of development, and to suggest that ownership and deep temporality are needed to resolve the problems. Chapter 3 will develop these themes further. It is sometimes thought that determinism is an entirely external problem, one that arises in a purely artificial elevation of abstract theory over concrete experience. I will argue that this is a mistake; experience can give rise to determinist intuitions, and do so from the first-person perspective. Determinism is then not only an abstract theoretical problem, but one internal to the very experience of agency. From here I proceed to develop a concept of deep temporality on the basis of Korsgaard's early work, ultimately concluding that it runs into the same aporia faced by other approaches to free will.

Here I change emphasis. Since attempting to develop a theory of free will runs into an aporia, I turn instead to an account of moral responsibility developed entirely without reliance on free will: attributionism. I approach attributionism by framing it as needed to offer a response to Galen Strawson's argument for the impossibility of moral responsibility, which plays in the chapters on responsibility a role analogous to that played by Nagel in the chapters on free will. I work out and defend attributionism as a response to Strawson in Chapter 4. Chapter 5 works out the control condition necessary for attribution of responsibility into a full-fledged Heideggerian theory of agency. Problems inherent in this account of agency and responsibility turn us again to ownership and temporality. In Chapter 6 I finally turn to the task of working out the Heideggerian account of full-fledged agency, complete with ownership and temporality, as inherently involving both freedom and responsibility.

2

Free Will

A. Eliminativism

The starting point for Nagel's conception of freedom lies in his distinction between the subjective and the objective standpoint or point of view or, to take another formulation, the internal and external perspective. Subjective and objective are, on this account, not two distinct standpoints, but operate instead in a continuum going from more subjective to ever more objective points of view. The movement from subjectivity to objectivity is described as a movement toward greater understanding:

> We can add to our knowledge of the world by accumulating information at a given level—by extensive observation from one standpoint. But we can raise our understanding to a new level only if we examine that relation between the world and ourselves which is responsible for our prior understanding, and form a new conception that includes a more detached understanding of ourselves, of the world, and of the interaction between them. (Nagel 1986 5)

Though Nagel does not cite the source from which this view seems to be adapted (namely, Hegel's *Phenomenology of Spirit*), the basic idea is a familiar one. It involves recognizing our own relation to the world as itself a part of the world open to investigation. Insofar as Nagel is committed to a naturalist perspective and thus a view that the "world" is approachable through largely empirical means, a certain tension immediately creeps up in this preliminary contrast between subjectivity and objectivity. On one hand, the transcendental position seems already to be excluded from any claims to objectivity, since objectivity necessarily involves taking oneself in all respects as

ultimately understandable within the world. On the other hand, however, the Kantian specter within Hegel's work materializes here: if we attain to greater objectivity by taking up a new relation to the world, one which now sees our previous relation to the world as part of that world, then objectivity has certain built-in limits.[1] A relation is always between two entities, but only *one* side of that relation (the world) is open to empirical analysis; the other side, the self, will always remain concealed. This tension becomes clearer in the following formulation:

> To acquire a more objective understanding of some aspect of life or the world, we step back from our initial view of it and form a new conception which has that view and its relation to the world as its object. In other words, we place ourselves in the world that is to be understood. The old view then comes to be regarded as an appearance, more subjective than the new view, and correctable or confirmable by reference to it. (1986 4)

This formulation is slightly misleading: the "ourselves" that we place in the world when we strive for a more objective view are not identical to the "ourselves" that examine this new world. Though something of the self must—as one of the relata—enter into a conception of the world that includes our prior relation to the world, something else must serve as a relata in this new relation. Insofar as any relation to the world, no matter how objective, is still a *relation*, something will be left out of it. And while some aspects of our subjective view may come to seem like mere appearance, that which stands outside of the objective view—that which relates to the world—simply cannot be taken to be appearance.

Nagel does seem to recognize the tension when he admits that, since we always see the world from a particular perspective, a fully objective view, a "view from nowhere," is not actually open to us. But the tension seems to have particularly

[1] See Žižek (1993) for a somewhat Kantian reading of Hegel along these lines.

problematic consequences in Nagel's account of freedom. As we will see, something very un-Hegelian happens in that account: when we take an objective view of our freedom, the self becomes integrated into the world and our understanding of the self changes; but our understanding of the world, oddly, remains the same. Nagel's point seems to be that it is the—in principle unattainable—ideal of objectivity that the subject be placed fully into the world, so that nothing remains on the subjective side of the relation. But this appeal to an ideal condition does not explain why the subject must be assimilated to the object. After all, a move toward greater objectivity might, for all Nagel has said, involve the assimilation of the world to the subject, and consequently a view that does not speak to every point of view from no point of view, but instead speaks from every point of view by speaking from one point of view. Nagel's account of objectivity is thus directed toward a certain outcome, a certain ideal of objectivity, but his conception of this ideal is biased from the start toward a subjectless world.

Nagel's account of freedom begins with a distinction between agency and action on one hand, and freedom on the other. Nagel suggests that agency is automatically threatened by a more objective view of the world: "*my doing* of an act—or the doing of an act by someone else—seems to disappear when we think of the world objectively." (1986 111) The solution, then, is to classify action as a basic and irreducible category. This does not, however, mean that freedom can be treated in the same way: we might have agency without having free agency. Clearly what Nagel wants here is to be able to analyze the problem of freedom in isolation from various difficulties in the philosophy of action, but he does not explain what he means in claiming both that action is irreducible and that such a view of agency is possible without a view of free agency. There is no

explanation of why freedom—which Nagel thinks is threatened by the objective view in the same way that agency seems to be—cannot likewise be seen as an irreducible category. On the other hand, many authors (e.g., Hornsby (2003 283), Velleman (1992 467)) have taken Nagel's account as raising the threat of disappearing agency, and not just disappearing freedom.

As is already clear, Nagel believes that the objective standpoint undermines our conception of ourselves as free, and his task in dealing with this is twofold. First, he must show in what way the objective standpoint *seems* to undermine freedom. Second, and more importantly, he must demonstrate that, when we attempt to discover what it is that the objective standpoint seems to undermine, "we end up with something that is either incomprehensible or clearly inadequate." (1986 113) This is an attempt to answer the former question of why we cannot save freedom from the threat of the objective view by declaring it, like agency, to be an irreducible category. The answer is that we cannot make freedom irreducible because we cannot explain what it is, though again there will be a question of how completely we can explain agency without reference to freedom.

Nagel approaches the problem of freedom in two parts, by breaking it up into aspects "corresponding to the two ways in which objectivity threatens ordinary assumptions about human freedom. I call one the problem of autonomy and the other the problem of responsibility; the first presents itself initially as a problem about our own freedom and the second as a problem about the freedom of others." (1986 111)[2] If both

[2] I should perhaps note that Nagel's use of the term "autonomy" here is problematic, since what he means by it actually matches up much more closely to what is generally meant by free will than by any accounts of autonomy. I will use Nagel's terminology, but with the caveat that his conception of autonomy, by insisting on the presence of having alternatives or being able to "do otherwise," in itself already contradicts the idea of autonomy. That may well strengthen Nagel's point; but it nevertheless seems odd to define a term in an idiosyncratic way simply in order to show that that term, on that definition, stands for something incoherent. To see the contradiction, we can look at Nagel's footnote on p. 116, in which, responding to

our own freedom and the freedom of others are threatened by the objective view, and if neither is intelligible, then the idea of freedom does indeed seem to collapse.

Initially it seems to us that we are constantly faced with open possibilities and that it is up to us to choose which of them to take up. But if we take an objective view, attempting to see ourselves from outside, we can consider not only the possibility as it appears to us—that is, as open—but also the various background conditions of our actions, both outside and inside ourselves. From an internal perspective, we see ourselves as having choices when we act. But when we take an external perspective, we must look not only at the choice directly behind an action, but also at the various antecedent conditions that make up the person, and then it seems like nothing is really open to us. "While we cannot fully occupy this perspective toward ourselves while acting, it seems possible that many of the alternatives that appear to lie open when viewed from an internal perspective would seem closed from this outer point of view." (1986 113)

Much rests on whether there is something coherent in this: if we fully lay out the various attitudes of the agent, "it is not clear how this would leave anything further for the agent to contribute to the outcome—anything that he could contribute as source, rather than merely as the scene of the outcome—the person whose act it is. If they are left open given everything about him, what does he have to do with the result?" (1986 113-114) Is there anything coherent in the idea of a "person" or, more specifically, an "agent," being merely "the scene of the outcome"? Isn't a "scene" in this sense itself not a person or agent? Freedom and agency seem to come bundled together, so that in making one

---

Wolf (1980), he expresses strong doubt about the idea that determination of any sort could be compatible with autonomy. But if autonomy involves the idea of self-rule, and self-rule is in turn understood as involving—at the least—consistency in our choices, then an autonomy without any determination whatsoever seems already to be ruled out. A fully indeterminate choice is arbitrary, and anarchy does not sit well with rule of any sort.

incoherent, we undermine the other as well. Furthermore, Nagel notes that the threat to freedom occurs regardless of whether antecedent conditions are seen as determining of actions. The question of whether or not we are "authors" of our actions seems to depend on whether we ourselves or various impersonal conditions are the causes of these actions.

When we view ourselves externally, we see ourselves as part of the world, rather than something standing apart from it. And once we see ourselves as part of the world, our freedom is threatened regardless of whether or not that world is deterministic. If this world is deterministic, then our actions are caused by antecedent circumstances, which are themselves caused by prior events, in a chain stretching back to the beginning of the universe. On this view, it does not make sense to say that *we* have anything to contribute. If, on the other hand, the world is not deterministic, it is still simply a collection of various entities and events, and these entities and events are still impersonal. While there may be genuinely open possibilities in front of us, the events involved in selecting one or another of these possibilities don't seem to be events of which we can claim to be authors. At this stage of the argument, Nagel is not trying to prove any point about freedom: he is only attempting to bring out an intuition, a doubt about whether or not we are free, or autonomous, that occurs to us the moment we start thinking about ourselves as merely parts of the world. As soon as we do that, we are bound to recognize that there are many things about our motivations that we do not know, and that these must cause even the motivations that we are aware of, and if the influence of all such factors is considered, the question is, what is left for *us* to do?

Replying to this doubt requires that we formulate an account of freedom or autonomy that can answer the question, and Nagel's main argument is that no such

account can be made coherent. The idea of autonomy, "presents itself initially as the belief that antecedent circumstances, including the condition of the agent, leave some of the things we will do undetermined: they are determined only by our choices, which are motivationally explicable but not themselves causally determined." (1986 114) Since the objective view explains the occurrence of every event in terms of causal mechanisms, the task of a theory of autonomy must be to provide a non-causal account of explanation. That is, it must be able to explain actions solely in terms of the motives and reasons of the agent. And this, Nagel thinks, is the real stumbling block, because any such account wants both to provide an explanation of action and to reject the possibility of such an explanation.

We can see this last point clearly as soon as we try to offer a non-causal but complete explanation for why an agent chose to do one thing rather than another. The answer, presumably, will be in terms of the agent's reasons for choosing that action. But if the agent had instead made a different choice, then we would again have to appeal to the agent's reasons to explain *that* choice. But since at the moment of any choice the agent only has a fixed set of reasons, this implies that whatever he chooses, each choice can be explained by means of the very same reasons. If the same set of reasons can explain both why an agent did one thing and why—if he had so chosen—he could have done another, then clearly this set of reasons does not provide a complete explanation of the choice; something more must be involved. But to explain this something more we are forced either to invoke further reasons, which leads to an infinite regress, or we must turn to an explanation in terms of the agent's character or dispositions, which themselves must be explained in causal terms. There is, on Nagel's view, no way out of this predicament:

either invoking the agent's reasons cannot explain why the choice was made to do one thing *rather than another*, or the invocation explains the agent's choice but only by relying ultimately on circumstances outside the agent's control.

Nagel suggests that what gives the idea of autonomy its force in the first place is precisely the objective view. Since this view opens up the possibility of knowing ourselves from the outside, it also seems to suggest that, once we know our motivations, we should be able to evaluate them and choose whether or not to be influenced by them in our actions. But this aspiration is clearly made impossible by that same objective view. To be able to evaluate and to either endorse or reject *all* of our motivational states, we would have to somehow separate ourselves from those motivational states. But this idea is incoherent. What, after all, is a self separated from all of its reasons and motivations? On what grounds does it endorse or reject anything about itself? We can evoke such a self only if we are willing to fully give up on any account of explaining its actions, but then we have simply conceded that the entire defense of the idea of autonomy rests on an inexplicable intuition. Either the self is part of the objective world, in which case it is ultimately causally determined in its choices, or it is outside the world, in which case it is no longer coherently described as a self, since it is a self without reasons and without motivations. "We belong to a world we have not created and of which we are the products" (1986 119), and therefore we have no way of standing apart from it. We want to fully encompass ourselves and create ourselves from the ground up, and at the same time we have no position from which we could do this.[3]

[3] Nagel is invoking a classic problem that first raised its head in theology: we are dependent, finite beings and yet somehow capable of acting with free will. As I will argue, Heidegger takes up this exact problem and reformulates it as the basis of a solution to the problem of freedom rather than the conclusion of an argument against it.

Nagel's argument for the incoherence of responsibility is largely the same. "To praise or blame is not to judge merely that what has happened is a good or a bad thing, but to judge the person for having done it, in view of the circumstances under which it was done. The difficulty is to explain how this is possible—how we can do more than welcome or regret the event, or perhaps the psychology of the agent." (1986 120) It seems we cannot hold anyone responsible for actions because actions are just events of a certain kind, and it does not make sense to pass moral judgment on events. Instead, we must enter into the psychology of the agent, seeing the alternatives open to him from his perspective, and understanding his reasons for taking the praiseworthy or blameworthy alternative.

The problem should be obvious in light of the foregoing account of autonomy. As soon as we try to understand the alternatives from the agent's motivational standpoint, we are presented with two possibilities. Either the agent did what he did for no reason at all, in which case it would be difficult to understand how he could be responsible for the outcome, or the agent acted for reasons, but these reasons are only the products of something beyond them over which the agent had no control. "Either something other than the agent's reasons explains why he acted for the reasons he did, or nothing does. In either case the external standpoint sees the alternatives not as alternatives for the agent, but as alternatives for the *world*, which *involve* the agent. And the world, of course, is not an agent and cannot be held responsible." (1986 123) The case of responsibility turns out exactly like that of autonomy: we have no way of assigning the action to the self that performed it as his responsibility. The action is either an event that occurred for no reason at all, or it is one that occurred as a process in the world.

In light of this point, I want to again question whether agency can be understood on an objective view even if free agency cannot. If alternatives are really alternatives for the *world*, rather than for the agent, the agent contributes nothing either to the explanation or to the ontology of action—"agent" is only the name for a particular, bounded scene of events. At least some minimal conception of freedom seems necessary if we are to have some minimal conception of agency.

This is, more or less, the whole of the story Nagel gives about freedom. Although he recognizes that the internal view of our decisions cannot be done away with, he does not think that anything close to a satisfactory treatment of freedom has yet been proposed. Consequently, both autonomy and responsibility are, on the final analysis, grounded in incoherent accounts. Nagel does, however, suggest that the objective view can offer another kind of freedom, one that may not match the aspirations of acting from outside the world, but at least operates as a substitute for the desires that lead us to believe in freedom. This substitute lies in morality. The belief in freedom, Nagel suggests, is really an expression of human beings' desire "to be able to stand back from the motives and reasons and values that influence their choices, and submit to them only if they are acceptable." (1986 127) The satisfaction of *this* desire does not require us to step entirely outside of ourselves; it requires only that we be capable of discovering and being influenced by motives, reasons, and values that are themselves objective. In other words, the desire for freedom is really satisfied by objective norms. Morality is "a practical analogue of the epistemological hope for harmony with the world." (1986 127)

Morality opens up the possibility for a kind of freedom, and consequently we are free—though only in a limited sense—only once we have discovered objective reasons

for actions *and* allowed ourselves to be guided by them. Morality can replace autonomy, Nagel argues, because it gives us a way of satisfying a desire: the desire to have objective criteria from which to evaluate our motives. I have a sneaking suspicion that this is a bit of a theoretical trick: although most people do evaluate some of their motives to some extent, it is hard to imagine that there is some universal human desire to act on objective criteria, and that this is what the desire for freedom really comes to. To anticipate my argument somewhat, this hope for harmony can be seen not as a conscious desire, but as an underlying transcendental feature of human existence. It will then be possible to act *in light* of it without achieving it: that is, roughly put, we do not need to "achieve" the level of objective norms in order to be free. Rather, we can be free by virtue of acting in light of our relation to the future; our actions can themselves be guided—well or poorly—by that hope for (or anticipation of) harmony.

As noted, Nagel claims that the problems of autonomy and responsibility initially present themselves to us, respectively, as problems about our own freedom and the freedom of others. If we look at autonomy and responsibility in this way, then it does not seem too difficult to admit that these are simply appearances—because we do not act with full knowledge of every event involved in the production of our actions, it seems to us as if our actions are fully under our control; because we resent and want to punish others, we fool ourselves into thinking that they are fully responsible. But something interesting happens if we take the problem of freedom as, at a fundamental level, a problem about our own responsibility. If I have acted wrongly and know that I have acted wrongly, then it seems to me that I am responsible for having acted wrongly. Of course this feeling of responsibility may itself be an illusion, but developing a philosophical

theory to demonstrate that I was not in fact responsible will come off as a kind of rationalization and self-deception. Of course it may turn out that this is right—that, in fact, I am not responsible for anything I do. But this point has implications that differ from the implication of the claim that *others* are not responsible for what they do. For even if I am not responsible for what I do, in claiming that "I could not have done otherwise," I am offering an excuse. If I fail to follow norms that I accept as binding, the excuse has a necessary ring of self-deception about it: I know what I had to do, but I failed to do it because what I do is not up to me. There are cases in which this will be accurate, but for it to be accurate at a *global* level, we would have to entirely give up on ever taking responsibility for who we are and what we do. Our relation to ourselves would be characterized by self-deception. Avoiding this problem, on my view, is the primary aim of developing a theory of free will. And avoiding this problem is crucial if we take moral responsibility not as a problem about third-personal freedom, but as a first-personal concern: unless I *can* be genuinely responsible, even having access to objective morality will not by itself allow me to *follow* that morality. From an objective standpoint, either the world—or at least the region of it embodied in me—will take up the moral norms or it will not; but *I* do not have control over whether or not I allow myself to be guided by morality. The problem of responsibility as a first personal problem, then, is guided by the question of whether or not I have any choice about the morality of my actions.

Let us return now to freedom and the question of what kind of freedom the objective view might require us to reject. Every choice, as an event in the world, must be caused by previous events. In the case of actions, the previous events are motives or

attitudes. For an action to be freely chosen, on the account Nagel lays out, its causes must also be freely chosen, since otherwise the action will depend on something outside the agent's control. The motives or attitudes themselves, then, would also have to be, or be the products of, choices. The same would go for any causal predecessors to the relevant motives or attitudes. But the string of causal choices will have to go back to before the agent's birth, all the way to the origin of the world, and certainly no agent is responsible for that. If any free actions were to exist, then, there would have to be at least one choice of the agent, making up some attitude (motive, reason, or value) that breaks this causal chain. But since this choice would be an event in time, it must have causal antecedents; the notion of a free choice is, therefore, incoherent.

What Nagel's account demonstrates, then, is the incoherence of a temporally shallow account of agency. Freedom, seen as an agent's ability to create in time an event without causal conditions, cannot make sense within a world in which an event, by definition, has a cause. A choice, as an occurrence in some present moment, must have prior occurrences causing it. It is therefore not surprising that, in offering a substitute for freedom, Nagel looks to "hope" for "harmony with the world." This would be no more coherent if this hope were for some future event that takes place outside the causal framework, but Nagel presents it instead as a hope for a kind of state, one in which the presence of causal antecedents to our actions becomes irrelevant. By extension, we might ask whether freedom is this kind of state, rather than property of an event. What Nagel rejects is free action; he does not touch essentially on free *will*, except insofar as the factors (motives, reasons, and values) that make up a will are themselves understood as actions. The problem with any temporally shallow account, in fact, is that it reduces any

candidate for freedom to the status of an event. But this automatically seems to undermine not simply freedom, but agency as such; if everything that occurs is seen merely as an event in a causal chain in the objective world, then it is unclear what actions have to do with an agent, except insofar as that agent is a scene of their occurrence.

What is happening here? In recognizing that the question of freedom is not likely to go away even after we take up the objective view, Nagel remarks that "I cannot say what would, if it were true, support our sense that our free actions originate within us. Yet the sense of an internal explanation persists—an explanation insulated from the external view which is complete in itself and renders illegitimate all further requests for explanation of my action as an event in the world." (1986 117) An action's originating "within us" is contrasted with that action's being an "event in the world." This contrast suggests that an action cannot originate both in us and in the world. The agent is already taken outside the framework of the world. And this removal of the agent as something internal from the world as something external creates a rift. Sealing that rift will require understanding how agency can operate within a world.

Nagel's argument attacks the coherence of the idea that an action can *originate* in us. Origination involves cause, and the argument is that neither an action nor its precursors can be caused both by us and by the world. But what we are trying to understand is what it means for something to be not merely an event in the world, but also *my* action, or *my* choice, or *my* attitude, and this semantically implies, first of all, not just origination but ownership. And ownership, obviously, does not imply origination or creation ex-nihilo. If we wish to understand agency, as well as freedom, we must turn to this idea of ownership. This, in turn, can help us to understand how an agent that is fully

in the world can, at the same time, stand in a different sense "outside" the world, at least insofar as "world" is understood as involving nothing more than objective, causal, relations. For what distinguishes agents from other entities, what places them outside the domain of mere causal explanation, and what accounts for "our sense that our free actions originate within us" is that we are the only sorts of beings who are capable of *owning* our actions, choices, and attitudes. Origination and authorship seem to be grounded in such ownership, not the other way around. To anticipate: that my will is *mine*, in some sense, is what stops the causal regress.

B. Compatibilism

Instead of taking up an overview of compatibilism as a whole, I will focus on the work of Donald Davidson and Harry Frankfurt, the two figures whose approaches continue today to exercise a major, possibly dominant, influence on work in philosophy of action and free will. This is largely because these two introduced analyses that describe action and freedom in terms of the relations between events or entities functioning at different levels. In Davidson's case, the important distinction is between actions (and intentions) and reasons; in Frankfurt's case it is between first-order desires and second-order volitions. These attempts open a path toward a temporally deep account of agency by structurally separating actions or decisions from their conditions. The strength of these compatibilist theories is that the tools of their analysis can be—and have been—put to work both by soft determinists and libertarians. Though Davidson is certainly no libertarian, Frankfurt expressly stated his intention to maintain an account

that *any* theory of agency, regardless of its commitments to causal determinism, would have to satisfy. But these theories also suffer from a corresponding weakness: their dual applicability makes them inherently unstable, and when pushed from either direction they tend to collapse into either eliminativism or libertarianism. This section, then, sets up the account of agency that is taken up by libertarians but, at the same time, is meant to work out what makes agency possible in a world of events.

Upon their introduction, however, these theories were unquestionably important, because they allowed for a middle path between the highly problematic compatibilism of the times and the mysterious libertarianism that seemed like the only alternative. The dominant form of compatibilism, which wanted to accommodate itself fully to the sciences and avoid metaphysical speculation, attempted to present freedom as simply a lack of constraint or compulsion, such as brainwashing or a mental disorder. A free action, on that view, is nothing mysterious: it can be easily explained by reference to the causal framework. On Ayer's account, for example, the apparently problematic nature of causal determinism arises entirely out of our loose understanding of this term. We tend to imagine an effect (an action) being somehow compelled by its cause, while "the fact is simply that when an event of one type occurs, an event of another type occurs also, in a certain temporal or spatio-temporal relation to the first. The rest is only metaphor. And it is because of the metaphor, and not because of the fact, that we come to think that there is an antithesis between causality and freedom." (Ayer 1982 22) But it is not so simple to decisively distinguish between compulsion and causation; if all we are doing is tracing out causal connections between events, there seems to be little reason to care about the types of events involved. On another note, as Frankfurt argues, this account only explains

free action, in a sense in which we can attribute it to any being with purposes from chimpanzee to spider. (Frankfurt 1982 82) If you hold down a spider's legs, it cannot move, but if you leave it be, then it seems to enjoy the same freedom of action as any human being. To distinguish between ourselves and spiders, Frankfurt proposes that another kind of freedom is needed: a freedom not merely of action, but of will.

The libertarian tradition attempted to break this reduction by introducing a different type of cause: unlike the causality involved in scientific explanation of events, occurrent causality, free action had to be explained in terms of agent causality. Normally, an event is caused by another event, which itself follows from prior events according to natural laws. A free action, on the other hand, is caused by an agent, and this cause has no necessitating precursors. The theory, as expounded by Chisholm (1982, 1966) and Taylor (1966, 1983), states that in order for an agent to be responsible for an action, two conditions must be met. First, the agent must be the sole cause of his action. Second, to avoid the action being randomly produced, the agent must have a reason for his action, stated in terms of purposes. Asking why an agent did something, "which is clearly a request for a reason, is almost never a request for a recital of causes. It is rather a request for a statement of purpose or aim." (Taylor 1966 141) But immediate problems with this theory arise. First, if an action is fully explained by means of reasons, it is unclear what the first condition—agent causation—really contributes to the explanation. Causal explanation is supposed to explain the action insofar as causes are thought necessary to explain events in general. However, the idea of an uncaused cause within the natural world, as an explanatory feature, has failed to achieve widespread support. If the appeal to agent-causality fails to explain anything, then perhaps we should simply stick to

occurrent causality. If, on the other hand, accounts in terms of reasons do not involve causes, we seem to be stuck with two incompatible sorts of sufficient explanations.

I. Davidson

This brief survey of the difficulties should, hopefully, suffice to show the importance of Davidson's seemingly modest proposal that reasons simply *are* causes of actions. It is the bringing together of these two conditions for action into a unified explanation that accounts for the historical importance of Davidson's major entry onto the scene with his "Actions, Reasons, and Causes." Davidson's causal theory of action attacked the orthodox Wittgensteinian view that explanations in terms of reasons or intentions on one hand, and those in terms of causes on the other, fall into two distinct language games that must not be confused. Davidson's response was to separate linguistic description from ontology: we can give two different descriptions of an action, and on one it will be rational (and intentional), while on the other it will be merely physical. But these explanations are compatible with each other because the action in question is, ontologically, a single event, where events are understood as "unrepeatable, dated individuals". (Davidson 1980f 209) This move will underlie the description of reasons in terms of causes.

The argument itself is typically elegant. Davidson argues that intention is a basic concept that allows us to distinguish actions from the broader class of physical events. He distinguishes between three sorts of situations in which an event (his example is the spilling of coffee) is attributed to an agent: "in the first, I do it intentionally; in the second

I do not do it intentionally but it is my action (I thought it was tea); in the third it is not my action at all (you jiggle my hand)." (1980b 45) The second case is the problematic one: I am holding a cup of coffee, but I believe it is tea. I want to spill tea and thus tip my cup, but to my surprise it is coffee that I spill. In this case my action of spilling coffee is not intentional; so how is it an action at all? Davidson's solution is to note that an action can admit of a number of descriptions. "My proposal might then be put: a person is the agent of an event if and only if there is a description of what he did that makes true a sentence that says he did it intentionally." (1980b 46) The spilling of the coffee in the second case is an action because under another description, like "trying to spill tea," it is intentional. An action, then, is an event that has at least one correct description on which it is intentional.

An intentional action, in turn, is an action that is rationalized by, or performed on the basis of, what Davidson calls a primary reason. "*R* is a primary reason why an agent performed the action *A* under the description *d* only if *R* consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property." (1980a 5) A primary reason is thus analyzed into two component parts. The first is a pro attitude, or an attitude in favor of certain types of actions. Davidson presents these attitudes in a very broad way, so that they contain all sorts of wants, desires, obligations, customs, values, and in general any sort of attitude capable of motivating an agent. The second component is a belief that a particular action that the agent has an opportunity to perform is an action of a type for which the agent has a pro attitude. Thus, if I jump out of an airplane, my action could be explained by my

longstanding (or sudden) need for a new thrill, coupled with a belief that jumping out of an airplane would provide a new thrill.

The task now is to show that these reasons are causes of the actions they rationalize. Davidson argues that unless we take reasons to be causes, we have no way of describing an action as intentional. To see something as intentional at all, we must be capable of interpreting it as rational. A piece of behavior of which we cannot, try as we might, make any rational sense is bound to appear as nothing but a reflex, a spasm, or a seizure. Being explicable in terms of reasons is thus a condition of possibility for being understood as an intentional action. But even if we grant that an action must allow for a rationalizing explanation, this does not yet guarantee that the reason for the action is a cause. Davidson's response here is that we have no other way of explaining the relation of the reason for the action to the action itself. Consider a case where an agent has two reasons for doing something: he might raise his hand to ask a question, or he might do so to stretch his shoulders. If he raises his hand and this is an action, then it must be done for one of the reasons. But we cannot appeal to the mere presence of either reason to explain the "because" in "he raised his arm because he wanted to stretch out his shoulder" in a non-circular way. We can only say that this is the reason because it is the reason, which fails to contribute any kind of explanation at all where two or more possible reasons are in play. Davidson's account allows us to explain what makes this reason into *the* reason why the action was done: the agent's desire to ask a question, together with his belief that he could do so only if he first raised his arm, *caused* him to raise his arm, and thus we can say that he raised his arm *because* he wanted to ask a question.

Action is thus explained entirely within a causal nexus; the agent seemingly plays no role whatsoever, leading to "an agentless semantics of action" (Ricoeur 1992 87)[4] in which an action is nothing but an event produced by other events. This may seem surprising: pro attitudes, after all are specifically described by Davidson as covering "not only permanent character traits that show themselves in a lifetime of behaviour, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action, like a sudden desire to touch a woman's elbow." (1980a 4) Are all of these events? Yes, it turns out. In the case of the latter sorts of attitudes, those which appear suddenly, it is not the attitude itself, but its onset that is the event under consideration. In the case of longstanding or permanent attitudes, on the other hand, the relevant event might be not the onset of the attitude itself, but rather the sudden recognition or realization that one has this attitude. My love of chocolate is not usually in the forefront of my thoughts, but springs into existence, so to speak, the moment I am face to face with a chocolate mousse. It is thus my love for eating chocolate, coupled with a belief that the cake in front of me is made of the stuff, that causes me to dive into it headlong.

Explaining just how reasons can be causes opens the way for Davidson's first strategy of getting freedom into the equation, which involves his theory of anomalous monism. The approach is designed to show, on the one hand, that we cannot construct any strict laws by means of which to predict mental events, due to "the holistic character of the cognitive field. Any effort at increasing the accuracy and power of a theory of behaviour forces us to bring more and more of the whole system of the agent's beliefs and motives directly into account. But in inferring this system from the evidence, we

[4] This is a polemical claim: the agent does play some role in Davidson's account, though Davidson never clarifies this role. Ricoeur's description is therefore apt, if slightly misleading.

necessarily impose conditions of coherence, rationality and consistency." (1980g 231) But this structure does not have the form of a physical system, and we thus cannot hope to find a direct correlation between the physical and the mental. This means that, even if we are in possession of all strict laws for predicting physical events, mental events will necessarily resist assimilation into this framework.

The theory is characterized by monism because Davidson does believe that all events are physical events, but it is anomalous because mental events are not subject to the laws that govern physical events. The reason is that a mental event is a physical event under another description. Since laws apply only to the events under their physical descriptions, the same laws do not apply to mental events. Furthermore, for Davidson, knowing all the physical events that occur would not tell us all the mental events—the correlation between mental and physical events is token-token, not type-type, which prevents any crude reductionism. Thus, for Davidson, although we can speak of actions as caused by reasons (because reasons, as mental events, are also describable as physical events), this also means that we cannot understand action—described as intentional, or in its mental aspect—as nomologically determined.

Whatever freedom this opens, however, is far more vague than Davidson suggests. For one thing, it is obvious that what is normally understood as freedom is missing—every action *is* explicable in terms of strictly natural laws; it just isn't explicable in terms of such laws *as* an action. Anomalous monism suggests that physical description cannot eliminate mental description, but this doesn't guarantee freedom of any sort; it guarantees only talk of freedom. Second, whatever is free here, it is not an agent. It is some set of events, perhaps a rationally and coherently organized set, but what

this set has to do with an agent is never specified. A real account of freedom would have to take the extra step of relating desires, intentions, reasons, and so on to the agent whose mental states they are; this is a bit difficult, however, given that all the states have been characterized as events.

Davidson's second account of freedom is more clearly pertinent to the standard discourse, and it is similarly compatibilist. What he wishes to show is not that freedom is compatible with determinism (he believes that "Hobbes, Locke, Hume, Moore, Schlick, Ayer, Stevenson, and a host of others have done what can be done, or ought ever to have been needed, to remove the confusions that can make determinism seem to frustrate freedom"[5] (1980c 63)), but that freedom can be understood as a causal power. This involves explaining how an action can be both free and caused. Typically, for Davidson, he approaches the question not by showing that freedom *is* a causality, but by attacking the leading objection to that view. The objection is that—as we have already seen—if an action is caused by something else, then the question arises whether or not the cause itself is free.

> In order to be eligible as a cause, the event mentioned must be separate from the action; but if it is separate from the action, there is, it seems, always the possibility of asking about *it*, whether the agent is free to do it… The only hope for the causal analysis is to find states or events which are causal conditions of intentional actions, but which are not themselves actions or events about which the question whether the agent can perform them can intelligibly be raised. (1980c 72)

The candidates for this position are the pro attitudes and beliefs of the agent. These conditions cause intentional actions but, at the same time, they are not actions themselves. "The antecedent condition does not mention something that is an action, so the question whether the agent can [or is free to] do it does not arise." (1980c 73)

[5] Anomalous monism, on the other hand, does seem to be an attempt to answer the first question.

At one level, this is a neat solution to the problem: freedom is a causal power because to perform a free action is simply for that action to be caused by an agent's reasons. The action itself is free because there is no further question about whether or not the agent was free to take up those reasons—the reasons are not things the agent does. Of course this is likely to be unsatisfying to someone who is actually interested in a theory of freedom, and Davidson notes this by mentioning that he does "not want to suggest that the nature of an agent's beliefs and desires, and the question how he acquired them, are irrelevant to questions of how free he, or his actions, are. But these questions are on a different and more sophisticated level from that of our present discussion." (1980c 73) Davidson does not return to these questions, though given his list of compatibilist heroes, one may assume that by the nature of an agent's reasons and how he came to acquire them, Davidson means something like whether the reasons are products of brainwashing, compulsion, or ignorance. But even given this caveat, we may ask whether Davidson's schema is satisfactory on the question of freedom.

Two pages later, he addresses the problem of overdetermination—that is, the question of whether it makes sense to say that someone's actions are free in a situation where he would have done the same thing even if he had not chosen to. It turns out that the answer is yes, because "even in the overdetermined cases, something rests with the agent. Not, as it happens, *what* he does (when described in a way that leaves open whether it was intentional), but whether he does it intentionally. His action, in the sense in which action depends on intentionality, occurs or not as he wills; what he does, in the broader sense, may occur whether or not he wills it." (1980c 75) Given that an agent is not physically prevented from doing something, and his doing of it is caused by his

reasons, the action is free. Otherwise, it is not even an action. But what it means to say that "something rests with the agent" remains mysterious. What makes the action free on this account is that it is (1) rationalized by a combination of a pro attitude and belief (a primary reason) and (2) caused by that reason. But surely this is where questions about (a) the nature and kind of acquisition of the pro attitudes and beliefs, and (b) the agent's role in *making* the reasons into a cause, come into play. The role of the agent cannot be irrelevant here: if there are no strict psychophysical laws under which the relation of reason to action falls (as anomalous monism maintains), then the causality of these reasons remains inexplicable. What Davidson's analysis shows us is that, given that an intentional action has been performed, we can then give a causal description of this action. It does not show how freedom could be a causal power—Davidson has only repeated that causal relations hold for all cases of acting. Davidson's account is thus internally incomplete. Something further is needed to explain what makes the reason cause the action.

A similar point arises with regard to Davidson's account of intention. In earlier essays, Davidson had looked only at two usages of "intention": the "intention with" and "intentionally." Both of these present intention as accompanying an action ("the intention with which Odysseus lied about his name," "Mercader intentionally stabbed Trotsky with the ice pick"); but what about "intending to" ("Heidegger intended to prepare the way for the fleeing or arrival of the gods" or "Bush intended to finish the war in Iraq within weeks")? The difficulty is that with the first two usages, intention is analyzed as a component of an action, but in the third case we have an intention directed toward the future, and such an intention may never be fulfilled by an action at all. For example, I

intend to write a novel but never get around to it; my intention may well be sincere, but other factors (laziness, lack of time, lack of talent) get in the way. What is this strange entity, this "pure intending"? After an extended analysis, Davidson finds a way to bring pure intendings in line with his overall theory without making them into private mental entities. He does this by distinguishing between prima facie judgments ("judgements that actions are desirable in so far as they have a certain attribute" (1980e 98)) and all-out judgments, or unconditional judgments that an action is worth doing. A prima facie judgment is merely a judgment, for example, that getting rid of eavesdroppers behind curtains is desirable. But such judgments are, despite providing reasons for acting, not sufficient to cause an action because it is entirely possible for me to have the judgment in question but also believe that the person eavesdropping behind the curtain is my greatest love, and I do not want to harm her. An unconditional judgment, on the other hand, is the judgment that a certain action is good all things considered; that is, in light of my other beliefs, it is the right thing to do. So, if I believe that the person behind the curtain is the king, then I might make the unconditional judgment to run the curtain through with my sword.

Somewhat incredibly and all too neatly, then, it turns out that a pure intention is really a kind of judgment that is identical to the judgment that accompanies any action. This is, perhaps, not so odd: since action is by definition intentional, it might seem to follow that there should be no important difference between the intention that accompanies an action and one that precedes it. The only difference is, for Davidson, a fairly minor one: a pure intention "is directed to the future." (1980e 98) In other words, it differs from the intention that accompanies an action only in that it involves a deferral of

the action (I do not say that the difference is only some stretch of time between the intention and the action, since the action might never occur—it is the deferral that is important). Ricoeur's critique comes in at this point. He argues that "Davidson has underestimated the unsettling effect that this addition of all-out judgment imposes on the earlier analysis" (1992 82), because this introduces a temporal dimension into the framework, and it is a dimension that necessarily involves the agent. Davidson has, essentially, interpreted pure intentions as a slightly modified version of intentions that accompany actions; but the analysis should go the other way: "with the delay there appears not only the character of anticipation—the intention's empty sighting, as one would say in a Husserlian perspective—but also the prospective character of the very condition of agency, as one would say in a Heideggerian perspective." (1992 82) In other words, we should interpret all intention in terms of intention to, reversing Davidson's analysis, because action as such involves as its condition a reference to the future. This reference, further, is inseparable from a reference to an agent, that is, the being for whom directedness to the future is a defining feature of action.

We can go a step further here: the elimination of the agent from the analysis obscures the relation between the intention and the action. If the action fulfills the intention's empty sighting, surely the agent is the one who brings this fulfillment into being, but the question of how is left open. If I have a pure intention and then act on it, what leads me from the pure intending to the action? Is it merely a further belief? Or a further judgment? A favorable set of circumstances in the world? We still need an account of the reason an agent acts at just that moment, which is not reducible to the reason for the action itself, which is just the reason for the intention. This question is

related to the earlier one of what makes a reason into the cause of an action. And by way of pure intending, it becomes tied also to the question of what connects any intention to the event that it makes into an action, or any reason to the action it causes. An agent is not merely a medium for these strange connections; not a "scene" in which the events of the world play out. The agent—to be an agent—must be active in this process, and that the process does not make sense without this activity lends credence to that idea.[6]

I want to wrap up the discussion of Davidson by tying it back to the line I have been slowly working out. What Davidson's account opens up is the possibility of a deep temporality: freedom is now made into a relation between attitudes and beliefs on one hand, and action on the other. The action may involve a choice about which it makes sense to ask whether or not, and in what sense, the choice was free. But the underlying states that cause the action are not themselves actions, they are not products of choice, and about them it makes no sense to ask such questions. But this relation between something that is a choice and something that isn't, a pro attitude, is quickly undermined. Davidson's view of temporality is entirely shallow: "The antecedent condition (*A* has desires and beliefs that rationalize *x*) is prior to and separate from the action, and so is suited to be a cause (in this case, it is a state rather than an event—*but this could be changed along these lines: 'coming to have desires and beliefs that rationalize x'*)." (1980c 73; italics mine) The same approach that first allows for a fruitful distinction between the action and the underlying attitudes—that is, the causal analysis—also reduces those attitudes, for all intents and purposes, to events. Even if they are, overall,

[6] I want to stress, however, that Davidson does not explicitly exclude the agent: quite the contrary. The agent is needed for deliberation (2004 107) and it is the agent's deliberative process that leads to various problems of irrationality and so on. But Davidson never clarifies the role of the agent in this account, nor it is clear how any strong notion of agency can be reconciled with the rest of the analysis.

states, they play their role in the causal nexus *as* events. They are, to be sure, not choices, but they are still clearly marked out as mere events in the world. The account, then, remains temporally shallow, though at the same time a possibility has been opened.

II. Frankfurt

This possibility is carried further by Harry Frankfurt, who attempts to fix the hole in Davidson's account by introducing the concept of the person into the analysis. An account of intentional action is, for him, insufficient if we want to understand what differentiates persons from other beings. After all, "human beings are not alone in having desires and motives, or in making choices." (Frankfurt 1982 82) What is lacking in an analysis of the Davidsonian type is the notion of a will, which is not brought in simply through an analysis of intentional action. *That* analysis can, at best, give us a notion of what it means for someone to be free to act in light of his mental states, and "this notion does capture at least part of what is implicit in the idea of an agent who *acts* freely. It misses entirely, however, the peculiar content of the quite different idea of an agent whose *will* is free." (1982 90) Frankfurt approaches this issue through the introduction of two central concepts: that of endorsement (which relies on having volitions of the second order) and that of identification (by means of which an agent's volitions are, in some sense, his own).

Frankfurt identifies the will with the agent's "*effective* desire—one that moves (or will or would move) a person all the way to action." (1982 84) Though a person may have any number of desires or wants, it is generally the case that most of them are not

ultimately expressed in action. The term "will," then, is limited only to those desires that are or would be if circumstances allow. These are what Frankfurt calls first-order desires. The crucial aspect of his account, however, is the introduction of second-order volitions. While first-order desires concern what an agent wants to do, second-order desires are instead directed toward first-order desires. Second-order volitions are second-order desires to the effect that a particular first-order desire be effective, that is, that the first-order desire constitute the agent's will.

Frankfurt's major step is to declare that having second-order volitions is the distinguishing feature of persons as such.[7] The distinction is clarified by his famous contrast between a person and a wanton. A wanton has first-order desires, and may even have second-order desires insofar as they are simply desires to have certain desires which the wanton does not have. Furthermore, the wanton may be a perfectly rational being; he can deliberate how best to carry out his strongest desire, and when to do so. But unlike a person, the wanton "does not care about his will." (1982 86) The wanton is simply not concerned about which of his first-order desires should be the strongest. A person, on the other hand, has second-order volitions that are concerned specifically with this. Frankfurt clarifies the difference by comparing a drug addict who always happily pursues his addiction whenever possible to one who does this unwillingly. An unwilling addict is a person—he has a desire to continue taking drugs as well as a desire to stop, and while he always acts on the former desire, he wants the latter to be effective. Although he cannot choose which desire will determine what he does, he is nevertheless a person because it makes a difference to him which desire will be effective. "When a *person* acts, the desire

[7] Sometimes he emphasizes *having* second-order volitions, while at other times he suggests that the *ability* to have them is constitutive of personhood, but the central point is clear in any case.

by which he is moved is either the will he wants or a will he wants to be without. When a *wanton* acts, it is neither." (1982 89)

This capacity for forming second-order volitions is naturally tied to a concern for the freedom of one's will. If the will consists of effective first-order desires, whereas second-order volitions are concerned with the issue of which first-order desires are to be effective—that is, the issue of what the person's will should be—the possibility of a free will is opened in the relation between these two levels of desire. "Freedom of action is (roughly, at least) the freedom to do what one wants to do. Analogously, then, the statement that a person enjoys freedom of the will means (also roughly) that he is free to want what he wants to want. More precisely, it means that he is free to will what he wants to will, or to have the will he wants." (1982 90) Freedom of action and will are sharply separated on this account: one can have a free will without enjoying freedom of action, and clearly one can have freedom of action without even having the capacity for a free will. With this step, Frankfurt opens the way for later theories of agency and free will, according to which either the ability to endorse or to control one's effective desires is the crucial mark of agency.

There are, on Frankfurt's account, two major advantages to this way of viewing free will. First, it explains why we attribute free will to human beings but not to other animals. The target here is the tradition of accounts claiming that an action is free when the agent is its sole originator. As Frankfurt points out, there is no particularly good reason to think that human beings are capable of originating their actions while other animals are not, and this account of freedom—freedom of action but not of will—is

therefore insufficient.[8] Second, this theory explains why free will is desirable: "the enjoyment of a free will means the satisfaction of certain desires—desires of the second or of higher orders… The satisfactions at stake are those which accrue to a person of whom is may be said that his will is his own." (1982 92) We want free will because it involves the satisfaction of desires we have; if we lack it, then our second-order volitions are left unfulfilled. Perhaps even more importantly, Frankfurt here introduces the notion of ownership, which I hinted at in my reading of Nagel. An agent whose will is the will he wants to have *owns* that will; if his will is not the will he wants, then he experiences it as something foreign. Free will, then, involves having a will that one feels to be truly one's own.

Frankfurt adds two important points to this account of freedom of will. The first concerns identification, while the second concerns moral responsibility. Taking them in turn, we may note that the issue of identification is crucial, because it is intended to be Frankfurt's reply to the problem of infinite regress, which is the obvious objection to his theory. One might ask: why is free will a matter involving only first- and second-order desires? Don't we—in order to have genuine free will—also need to have the second-order desires we want? That is, don't we need our second-order desires to correspond to our third-order desires? But why stop there? Frankfurt admits that the possible number of orders is theoretically limitless, but attempts to dismiss the difficulty. The series of ever-rising orders is terminated, "when a person identifies himself *decisively* with one of his first-order desires". (1982 91) If this occurs, there is no further important question of whether the relevant second-order volition—the identification involved—is the desire the

[8] Adding reasons to this account, as Richard Taylor does, would not help, since a wanton may both originate his action and have reasons for them, and yet not be a person.

person wants to have from a higher vantage point: that answer is provided in the decisive identification itself, in the commitment made to the first-order desire.

Second, Frankfurt is eager to use this new theory of free will to bolster the claims he had made two years earlier concerning moral responsibility. There, the goal had been to undermine the principle of alternate possibilities, which "states that a person is morally responsible for what he has done only if he could have done otherwise." (1969 829) Essentially repeating the argument of the earlier piece, Frankfurt here distinguishes between having a free will and acting *of* one's free will. On his view, only the latter is required for moral responsibility. The distinction is drawn in the following way: "a person's will is free only if he is free to have the will he wants" (1982 94), but acting of one's free will requires only that the person's will is the will he wants to have. If the latter condition holds, then regardless of whether the agent *could* have acted otherwise, he would not have, because he acted the way he wanted, "and even supposing that he could have had a different will, he would not have wanted his will to differ from what it was." (1982 94) This latter condition is important because, given the agent's identification with his first-order desire, his will is his *own*; he cannot complain that it was outside of his control, because it is the will he wanted to act on. The further question of whether or not his will was free—that is, whether or not he really could have chosen a different will—is irrelevant to judgments of moral responsibility because even if he *could* have had a different will, he would not have chosen to. Free will is thus not needed for moral responsibility because it does not contribute anything; in these situations, it would not change how the agent acted. Frankfurt illustrates the point with an example that, regrettably, does not make his claim significantly easier to swallow: an unwilling addict

is not morally responsible for his actions, because he experiences his addiction as something foreign to him, as something outside his control; he feels he is being pushed around by external forces and is himself reduced to the role of an innocent bystander. But a willing addict, an addict who identifies with his desire to follow through on his addiction, takes his will as his own: he is therefore justly held responsible for it.

If something here does not seem quite right, it is because there are at least two major problems in Frankfurt's theory. The first concerns the notion of identification, and whether this is sufficient to stop the infinite regress while still maintaining free will.[9] Called on to explain the notion of identification, Frankfurt defines it as the agent's endorsement of a first-order desire (by means of a second-order volition), together with a satisfaction with that endorsement. Satisfaction, in turn, means that the agent is not interested in repudiating the second-order desire. (Frankfurt 1992) This account, unfortunately, is insufficient, since there are many possible reasons why someone might be satisfied in this way. In addition, although Frankfurt strives for a theory of free will that would be satisfactory to all sides, clearly this notion of identification cannot be enough for libertarians: if identification with some first-order desire is, ultimately, a matter of chance or, as Frankfurt argues in his later work, even a matter of necessity, then the entire structure of endorsement and satisfaction leads to nothing more than an illusion of free will. An agent who is satisfied with his endorsement may well feel that his will is free, but this feeling is only a self-deception, since ultimately he cannot help feeling that way. If the implication, then, is just that free will is a matter of feeling—one either feels

[9] The classic criticism is leveled by Gary Watson, who questions whether a desire of any sort is capable of making a will into one's own will: "We wanted to know what prevents wantonness with regard to one's higher-order volitions. What gives these volitions any special relation to 'oneself'? It is unhelpful to answer that one makes a 'decisive commitment', where this just means that an interminable ascent to higher orders is not going to be permitted. This *is* arbitrary." (1982 108)

one's will to be free or one doesn't, and that's the end of it—clearly something is still missing from the account.

What this something is might perhaps be gleamed from an inspection of the second problem: Frankfurt's vacillation between two notions of free will, a vacillation he exploits to great effect. We can catch a certain ambiguity already in the distinction between having a free will and acting of one's free will. As we have seen, Frankfurt claims that, so long as the agent wants to have the will on which he acts, he is responsible for his action because "he did it freely and of his own free will." (1982 94) But this implies that having a will that is one's own does not actually require free will at all—it requires only that the agent wants to have the will that he does. And this, in turn, puts into question Frankfurt's argument that his theory can explain why free will is desirable. He argues that it is desirable because it involves the satisfaction of a higher-order desire. But a morally responsible agent is an agent whose higher-order desires are satisfied and who may yet lack free will! What Frankfurt shows, then, is that moral responsibility is desirable, not that free will is.

The ambiguity in question is that between being able to have the will that one wants, and simply having a will that corresponds to the will one wants to have. Frankfurt himself is well aware of the distinction, since it underlies his argument against the principle of alternate possibilities. Furthermore, he points out that "it is in the discrepancy between his will and his second-order volitions, or in his awareness that their coincidence is not his own doing but only a happy chance, that a person who does not have this freedom feels its lack." (1982 90-91) In other words, someone whose first-order desires happen to correspond to his second-order volitions, yet who does not feel that he himself

has any power over this correspondence, lacks free will. Free will requires a level of agency such that, "with regard to any of his first-order desires, [a person] is free either to make that desire his will or to make some other first-order desire his will instead." (1982 94) Yet immediately after stating this criterion, Frankfurt points out that it is "a vexed question" of how this sort of freedom is possible, and affirms that it is not needed for moral responsibility. But that claim seems to be in direct conflict with the following account of the morally responsible agent who may not have free will: "since the will that moved him when he acted was his will *because* he wanted it to be, he cannot claim that his will was forced upon him or that he was a passive bystander to its constitution." (1982 94; italics mine)

This cannot be right. The problem is one of how the "because" is to be interpreted. If we interpret it to imply nothing more than correlation—simply that the agent's will happened to be the will he wanted—this will not be enough, following Frankfurt's own admission quoted in the previous paragraph, for the agent to avoid feeling that he was "a passive bystander": that is precisely how he *must* feel, if he recognizes that the correlation between his first- and second-order desires is not his own doing; and thus the correlation alone is not sufficient for a person's will to be his *own*. The "because" must—following Davidson—be seen as involving causation; the agent himself must somehow cause his will to conform to his second-order volition. Only in this way can his will be his own.

Frankfurt's attempt to separate the possession of free will from the conditions of moral responsibility therefore fails, and it fails according to the terms of his own account. An agent without a free will does not *own* his will. And the problem is exacerbated once

we realize that, per the previous analysis, even if the agent feels satisfied with his endorsement of his will, there is no clear reason to think that this alone provides a sufficient condition for ownership. The attempt at compatibilism either becomes eliminativist (there is an event of identifying with some first-order desire and an event of feeling satisfied with that endorsement) or it requires a libertarian account of freedom, according to which agency extends beyond identification. This latter follows from the even more serious problem raised by the foregoing: what Frankfurt has left out of his account of freedom of the will is, oddly enough, freedom of the will. He has, to be sure, given at least a preliminary definition of what such freedom entails, yet there is no account of how it might be possible, or of what exactly the crucial term—the term responsible for the "because"—might be.

Since, on Frankfurt's account, it is possible to have a correlation between first- and second-order desires without there being actual freedom involved, the implication is that desires or volitions are not of themselves active: the simple fact that I want my will to be *X*, and at the same time my will is *X*, does not in itself provide evidence that these two events are connected in any non-contingent way. The causal, or agential power involved cannot, then—at least on the implication of Frankfurt's analysis—be located in any feature or aspect intrinsic to desires. It is something else, something that only agency can bring to the fold. But surely *that* is the question of free will. Unless it is answered, a further infinite regress threatens to open. If the will is the dominant first-order desire, while free will involves the selection of a will that corresponds to a second-order volition, it seems that some will must be involved in connecting these two desires. But if there is,

in this way, a will behind the will (so to speak), what prevents us from having to admit that yet a further will might be needed to ensure the efficacy of this one?[10]

What I have been suggesting is that Frankfurt does set out to close a major gap in Davidson's theory: it is not enough for freedom that an action be caused in the right way by the desires (and beliefs) that rationalize it. A further structural level is needed, and this is provided by distinguishing between orders of precursors to an action—a desire and an agent's relation to that desire. The sort of freedom that matters, for Frankfurt, is not contained in the relation between reason and action; it involves the further relation between the reason and the person for whom it is a reason. What Frankfurt's account lacks, however, is precisely a way of making the leap that he wants to make: the leap to the person. In presenting the relation as one between desires of different orders, he gives us (with the addition of relevant beliefs) only a relation of reasons to the further reasons *for* those reasons. On this level, then, his account does not differ in any major way from Davidson's compatibilism, which certainly also allows that mental events like reasons be caused by further reasons.

But Frankfurt's account is also different in a crucial way, because he at least opens up the question of the person, or agent. He takes a deliberate step back from the action itself and asks about the relation of the person who performs this action to the primary reason that causes it. The notion of owning a desire or will, though ultimately given an unsatisfactory exposition, is a major addition to the philosophy of action. An action is rationalized by reasons, which are simply kinds of events. But those reasons are

[10] Watson's solution, which emphasizes an agent's values or *evaluations* rather than second-order *volitions*, may seem to help with the earlier problem of infinite regress in desires—though I do not think it ultimately does—but it is of no help here. For the question of what can cause values to be effective is just as problematic, if not more so, as the question of what makes desires effective.

endorsed by persons, who are not events. Frankfurt therefore deepens—or points to a way of deepening—the temporality implicit in a philosophy of action. New kinds of attitudes—identifications and endorsements—are introduced in relation to any choices, which are now relegated to the background. What takes center stage on a Frankfurtian theory is not the choice at all, but the agent's ownership of it. But to get us anywhere, the account requires a further explication of identification and ownership. To get past the Davidsonian theory of events causing each other, we need to establish that the identifications and endorsements are not themselves simply events.

As I mentioned earlier, compatibilist theories derive their strength from the attempt to give an account of freedom without regard to the question of determinism: although Davidson explicitly embraces it while Frankfurt insists on remaining agnostic, both lay out what they consider to be structural or relational features of agency, which any theory would have to account for. But we can clearly see how both writers could easily be pressed into an eliminativist mode. If (in Davidson's theory) we are merely dealing with events linked in a causal chain, and if (on Frankfurt's approach) what allows for freedom and responsibility is a certain identification, which may (for all we know) be necessitated by natural causes, then the resulting structures might explain the *feeling* of freedom, but they will not be sufficient to account for responsibility; freedom will be a mere surface phenomenon, even if (accepting anomalous monism) some version of it can never be fully eliminated from the account. But on the other hand, I have argued that Davidson and Frankfurt both share a common problem: the problem of accounting for the "because." Frankfurt leaves it mysterious; Davidson insists that we can explain it only through causality, but he does not account for the causality itself, or for the effectiveness

of the cause. A certain relation, apparently crucial to freedom, is contained in the "because," regardless of whether the statement in question is of the form "*A* did *X* because of *Y*" or "*A* was moved by will *X* because he wanted to be moved by will *X*." How, then, are we to make sense of the agency that seems necessary to account for this "because" in terms that do not simply eliminate freedom? It is in dealing with this question that libertarian accounts come into their own.

C. Libertarianism

Libertarianism is characterized by the insistence that free will requires the existence of indeterminism. This united rejection of determinism is, of course, an ancient one, but in contemporary debates it is largely a response to the Consequence Argument put forward by van Inwagen.[11] The argument essentially formalizes the basic intuition behind incompatibilism: if determinism is true, then what one does happens necessarily given the state of the universe at any prior time together with the laws of nature. But no one has or ever had any choice about the state of the universe prior to their birth, not to mention prior to the existence of life on earth. Thus, if determinism is true, no one has or ever had any choice about what actions they will take.

Some libertarians, like Goetz (1988) and Ginet (2007, 1990), simply reject the notion that causation of any kind is consistent with free action. They argue that the belief that free actions must be caused is simply misguided. But most recent libertarians have avoided this train of thought due to the difficulty of making sense of uncaused events.

[11] In, among other places, Peter van Inwagen, *An Essay on Free Will*. (Oxford: Clarendon Press, 1983) and "The Incompatibility of Free Will and Determinism," in *Free Will*.

First, such events seem to be out of place within a universe in which events are normally parts of causal chains; second, as already noted, the post-Davidsonian presumption is that explaining the relation between an action and the reason for which it was taken requires a causal account. Noncausalists have responses to these concerns, of course; but the responses have not been widely convincing. Most libertarians have therefore followed the roughly Davidsonian line that reasons and causes cannot be kept entirely isolated from each other. In their broad outlines, the theories attempt to respond to the question I raised against Davidson: the question of what role the agent plays in making a reason effective. In some sense, the agent's role is taken the be causal; the libertarian strategy differs from the compatibilist one primarily in the claim that the causation involved in action may be nondeterministic. Frankfurt is also taken up, though usually with some tweaks. Freedom of action is generally taken to involve a relation between the action-causing events (desires) and some underlying mechanism: an agent, a will, or a preference.

Ekstrom picks up a largely Frankfurtian account, arguing that, for the most part, compatibilism works: autonomous actions may well be determined, so long as they are determined by the agent's preferences. Agents themselves, on the other hand, are made up by attitudes coherently structured along the lines of those preferences together with acceptances, or endorsements of certain judgments about the good, so that "it is reasonable to conceive of *the self* as constituted by an aggregate of preference and acceptance states, along with a certain capacity" (2005 54), which itself consists of the ability to form and reform one's character through the choosing of preferences. So much of the discussion of which attitudes or beliefs are internal to the agent or external can

proceed along Frankfurtian lines. The major difference is that responsibility itself requires that the preferences be chosen nondeterministically. (2005 57, n. 33)

So long as a preference is "formed by a process of critical evaluation with respect to one's conception of the good." (2000 106), and its formation is not causally determined, it represents the agent. Thus, any action that takes place on the basis of such preferences—that is determined by such preferences—is itself free. But there is a concern here: since the agent is simply *identified* with her preferences, and these are the products of nondeterministic deliberation about the good, any usual sense of agency seems to be lost. Actions are free insofar as they are determined by preferences; the preferences result from deliberation. The agent is, once again, only a site for that deliberation. Now I do not think this is itself a problem; the difficulty is that the preferences are themselves chosen through *action*. But if action—in order to be produced by an agent—must take place on the ground of preferences, what we have are preferences choosing other preferences, and this seems unhelpful.

Randolph Clarke, a major recent exponent of the agent causal theory, argues that we can reconcile freedom with event causation so long as we allow that such causation is probabilistic rather than deterministic. As a result, although every action is caused by prior events, no set of these events—without the agent's intervention—is sufficient to cause the action. The nature of this intervention, in turn, takes the following form: "what an agent directly causes, when she acts with free will, is her acting on (or for) certain of her reasons rather than on others, and her acting for reasons ordered in a particular way by weight, importance, or significance as the reasons for which she performs her action." (1993 194) The Davidsonian notion of reasons as causes is preserved; the only point here

is the addition of an agent—something Davidson does not expressly exclude, but simply fails to explain. And Clarke fixes the difficulties inherent in earlier agent-causal theories—such as those of Chisholm (1966) and Taylor (1966, 1983)—by offering what he calls an integrated account, one on which the agent causes his actions in *conjunction* with other events.

Agent causation, on this view, "is (or involves) *exactly* the same relation as event causation. The only difference between the two kinds of causation concerns the types of entities related, not the relation." (1993 197) Clarke thus tries to demystify agent causation by making it an ordinary kind of causal relation, though the "agent that is a relatum of such a relation is not identical to any event, property, fact, or state of affairs, nor to any collection of such things." (1993 196-197) Much of Clarke's positive account aims at making sense of the notion of an agent as a substance and defending the coherence of substance causation in order to avoid letting it be reduced to event causation on the one hand, and appearing completely implausible on the other.

Let me take just one point: a standard objection has it that agent causation cannot explain why particular actions are performed at specific times. If agents are enduring substances, they are present as a whole at each moment in time; but why would a substance present as a whole at each moment act at one moment rather than another? Clarke responds that his integrated account allows for a solution: "whether a certain agent who possesses an agent-causal power is disposed to act at a given time, and, if so, which specific actions she is empowered to cause then, depend on factors such as which reasons or motivations the agent has then… A fact of this sort, it may be held, may explain why it did not happen yesterday but did happen today that she caused this particular action."

(2005 202) So we can explain why an action was performed at one time rather than another in the usual way: by appealing to other events—the presence of certain reasons or motives—that prompted the agent to cause the action.

Clarke recognizes that it is the reasons and motives that we are generally interested in when trying to explain an action, but he notes that agent causation isn't meant to contribute to action explanation (though he suggests that it might); rather, it is supposed to solve the problem of control, very roughly, the problem I have been raising of how the *agent* enters into his actions. Rather than reducing the agent to particular mental functions, events, or simply a stage on which events play themselves out, Clarke wants to insert the agent *as such* into the picture. But however this is supposed to work, it is not clear that postulating an agent really solves the problem. Pointing to something quite different from the other sorts of entities we encounter in the world and saying that it can help explain what role the *agent* plays in causing actions strikes many as unconvincing, despite Clarke's attempts to defuse the major criticisms.

Furthermore, let's say that the agent is a substance that acts in concert with certain events (desires, motives, etc.) to cause another event (an action). The agent is not reducible to any of these events; but, as substance, the agent is also not *constituted* by any of those events. If she were so constituted, this would simply be the kind of reduction Clarke is arguing against. But if the agent's motives, desires, choices, and action do not make up who *she* is, then we seem to lose much of what we might normally want in a theory of agency: what it is that makes us who we are. To say that we are just substances adds a formal feature to agency, but it doesn't help satisfy the major concern: how it is that *who I am*, as this particular agent, can cause my actions? What is individual to me

seems to belong to the events that jointly cause my action. Thus, in attempting to stop the regress involved in tracing the causes of actions back to further events, themselves the products of previous events, agent causation seems to lose track of the agent's role in the process after all.

I am not sure that Robert Kane's influential event-causal account fares significantly better. The account itself, however, is quite subtle, and once again I will not do it justice here. It involves taking up a roughly Frankfurtian model, but building in the condition that the agent must be the ultimate source of his actions: "*free will* – as opposed to mere *freedom of action* – is about the forming and shaping of character and motives which are the *sources* or *origins* of praiseworthy or blameworthy, virtuous or vicious, actions. Free *will* (in contrast to mere free *action*) is about *self-formation*." (2007 16) Essentially, Kane breaks down what he calls the Ultimate Responsibility condition into three aspects, which I here oversimplify: (1) the action must be produced by the agent's effort of will, (2) the action is rational,[12] and (3) the conjunction of (1) and (2) provides the complete and only explanation of why the action occurred. In other words, it is important, "first, that the agent *produces* or *causes* the outcome, and second that the agent's doing so is rational." (1989 232) Kane's concern is with the standard anti-libertarian argument, i.e., that indeterminacy alone is not sufficient for freedom. That indeterminacy must be incorporated into the account in such a way as to explain the action as dependent on the agent and as rational. Thus, the above is meant to satisfy two conditions, the Ultimacy Condition (which places the agent at the origin of the action by eliminating non-agential sufficient causes) and the Explanation Condition (which

[12] "The agent (r_1) has reasons for doing so (whichever occurs), (r_2) does it *for* those reasons, (r_3) does not choose (for those reasons compulsively), and (r_4) believes at the time of choice that the reasons for which it is made are in some sense the weightier reasons." (Kane 1989 232)

answers the question of why the agent did what he did and not something else—the point here being to prevent the arbitrariness that might seem to be associated with indeterminacy).

There are two points in Kane's analysis that I want to call attention to. First is the reference to the effort of will. The agent causal account, avoids the problems raised by placing some extra action—like an effort of will—in front of the action in question by linking the agent causally to an action performed for certain reasons. Kane deals with the problem by similarly insisting that we need not place another action of some sort before the effort of will. The effort is the result of certain conditions, like the character and prior motives of the agent, which raise the question of how the agent is to act in the first place. But these conditions do not *determine* the outcome of the deliberation: the effort of will involved in choosing a course of action is itself the location of indeterminacy; starting with a set of conditions, it arrives at a conclusion of how to act given these conditions.

Second, there is the problem of rationality: as stipulated, the agent's final decision must be seen by him as rational. But the indeterminacy requires that the agent could have decided otherwise, and that decision, too, would be seen as rational. Furthermore, either decision would be seen as rational by the agent given exactly the same set of background conditions. This problem threatens to turn any libertarian theory into either a theory of complete determination by reasons, or a theory of irrational decisionism. Kane's answer is subtle, but not essentially dissimilar to Clarke's: "in libertarian choice we must be choosing the reasons that in turn explain our action." (Kane 1989 246) The account is largely Davidsonian: an action appears rational to an agent in light of the reasons he has (or at least the ones he considers in the deliberation) and the weight he assigns to them.

The further step here is that the choice itself determines the relative weight of those reasons for the agent. Whether the agent chooses to do one thing or another, then, he will have reasons for that choice, and these reasons will seem like the better reasons for action because the agent has chosen them, in the same deliberative process, as weightier reasons. Though Kane admits that this solution is circular, he does not think it is empty—it leaves something missing, but it explains everything that needs explaining. What is missing, or what seems to be missing, is just an account that would make the final decision *necessary* in light of the prior conditions; but that, of course, is precisely the point of a libertarian theory. We may also note that what makes the final decision into the agent's own is not that he somehow chooses the underlying motives, but that he decides what respective *weight* to give them in producing his action. This, despite some modifications, is essentially the Frankfurtian account of identification: the emphasis is placed on choosing among opposing motives, which explain both why the agent makes an effort to choose and (since there is an opposing motive) why this is an *effort*. The difference from Frankfurt is just that the deliberation that issues in choice is indeterministic. (Kane 1996 128)

Like Clarke, Kane argues that absolute freedom—in the sense of being the ultimate origin of one's action—is required for moral responsibility, and he insists that such freedom is incompatible with determinism:

> The Epicurians held that if there was to be room in nature for human freedom, the atoms must sometimes 'swerve' from their determined pathways. If the atoms do not 'swerve'—if the appropriate 'causal gaps' are not there in nature—then there is no room for an incompatibilist free will in nature. One would need a Kantian noumenal order, or some similar stratagem, to make sense of it, and this I think should make incompatibilists uncomfortable. (1989 231)

Kane's "comfortable" solution appeals to quantum indeterminacy, essentially the last place in which one might still hope to find indeterminacy in nature, which on his view may be produced by the effort involved in making a choice. This sounds highly unlikely—and some have argued that, based on what we already know about the brain, it is false. But interestingly Kane is not willing to give up on freedom however the empirical science turns out—naturalism is only the (currently) most plausible way of attaining it. We need freedom, because without it we cannot have moral responsibility. Kane's wording clearly implies that even if the gap in nature turns out not to be there, he will still maintain that we are free—we'll just have to find a different, more metaphysically troublesome, way of explaining this!

Kane sticks to the quantum explanation because of his naturalist convictions, but there is a further point to note here. The "further stratagem" that Kane finds so "uncomfortable" would, on his view, involve not some change in our conception of the agent, but rather a different conception of the agent's role in the world of natural causality; in other words, even if a naturalist account of freedom should fail, we must still hold on, at any expense, to a fully naturalist conception of agency. Although the implication seems to be that libertarian freedom is somehow essential to our understanding of agency, Kane is still content with understanding the agent in essentially the same way as the compatibilists. True, he does add a notion of indeterminate willing, but this does not change the agency—it changes only the nature of the agent's willing and, even there, it says nothing especially new about the willing itself, but only adds a component—indeterminacy—to it. This, I think, is a crucial error. If we take freedom to be not a feature of agency as such, but a property superimposed on agency, then the

libertarian and the compatibilist account will simply converge. *This* should worry libertarians far more than the discomfort of having to develop a new account of agency. What I have been arguing is that Kane's account is ambiguous. On one hand, the concern to preserve freedom in order to allow for responsibility seems to suggest that responsibility is somehow internal to our conception of agency itself. On the other hand, if responsibility is linked to freedom, and freedom is not an internal feature of agency, then it is unclear why we cannot happily maintain our view of ourselves as agents without invoking indeterminism. Compatibilists *are* happy to maintain that the freedom allowed within determinist constraints is perfectly sufficient for all the responsibility we need. If we accept their notion of agency, why not accept the rest of the account?

Even more problematic, however, is the narrowing of the domain of freedom implicit in theories of this sort. Free will is defined exclusively in terms of indeterminacy plus rationality. But this reduces free will to free action. Kane, as we saw, emphasizes freedom of *will*, to the extent that he calls his account "free willist." His view, then, is that agents must be the ultimate sources of the purposes on which they act; that is, they must be the sole choosers of what constitutes their will through what Kane calls self-forming willings. But self-forming willings involve actions of, at some point, choosing the motives that will heretofore determine how we act. Kane's claim is that all self-forming actions are self-forming willings (1996 125). But the reverse also holds: we form our will through action.

Thus the question we saw raised by Frankfurt's account—of how the will itself can be free—falls by the wayside: the will is free only because it is formed by free actions. But the question about free action is, arguably, not the right sort of question. Let

us admit, for the moment, the possibility that there is a genuine indeterminacy in our decision making. What could possibly rest on that? The idea, for both Clarke and Kane, seems to be that responsibility involves origination, or production: we have to establish a causality on which the agent alone determines himself to action. But this gives rise to a problem we have already seen: the agent is not responsible at all for the pre-conditions of his action. His responsibility occurs only in the deliberation between those preconditions and the action itself. But this makes it look like the action, though undetermined, is now entirely random, and although we can perhaps assign responsibility for it in the sense of origination, we do not get to moral responsibility. The solution to this problem is supposedly provided by the heavy reliance on reasons: I select based on reasons. But this doesn't help.

There are two points here. First, my selection among reasons itself seems random, and then we are back to the problem of randomness. Kane's argument that the act of choosing both establishes which reason is most weighty and provides a rational justification for taking that reason as most weighty could only help us in explaining the choice—whichever way I choose, my decision can be explained in terms of my reasons. But this doesn't seem to allow for responsibility any more than irrational randomness. Second, the pre-conditions of my choice—my motives, reasons, desires, values—are all there *before* the deliberative process begins. The materials on which I deliberate, on the libertarian account, are pre-given; the best I can do is select among them. I agree that this is *necessary* for responsibility, but it is not sufficient to establish responsibility: it is possible, for example, that all the pre-conditions I am considering are bad ones. In that case, even though I originate the decision reached on their basis, I still cannot be held

responsible for it because I am not responsible for the framework within which the choice is made. That is, libertarian theories that focus on indeterminacy fail to answer Nagel's challenge: how I can be responsible for anything I do given that I am not responsible for (at least some of) the preconditions of my action. Indeterminacy in my deliberation, even coupled with rationality, is simply not sufficient for responsibility.

There are attempts to address this problem, of course. Both Kane and van Inwagen (1989) attempt to show, though with a slightly different goal, that the scope of our responsibility is broader than the scope of our free action. Invoking the Aristotelian account that our actions habituate us into patterns of action, they argue that our free actions result in our acquiring certain character traits—if I freely choose to lie, for example, then I become accustomed to lying so that it does not seem to me that lying is especially wrong; but since this habit of lying is the consequence of my free choice, I am responsible for it. There are, some obvious problems with this account, of course: for example, we lack knowledge of the effect to which our actions will influence our character, and in the absence of such knowledge, it is difficult to assign responsibility—after all, even if we freely choose the action, we certainly do not freely choose all the unknown effects of that action. But this also doesn't expand the field of responsibility far enough: my free actions do not occur starting with my birth; in fact, for Kane and van Inwagen, because of the stringent conditions imposed on the definition of free action, it turns out that very few of our actions really are free. But this implies also that the vast majority of my character traits, motives, and values are completely outside the scope of anything I did or could have done. But since *those* features, for which I cannot conceivably be held responsible on this account, will likely play a role in even my free—

or undetermined—actions, it is still unclear how responsibility can be assigned to those actions. My suggestion, then, is that something is missing in this account of responsibility. No set of conditions, added to the conception of agency taken up by libertarians, is likely to fill the hole. What is missing, I believe, is a conception of agency that contains responsibility as an internal feature: either we are the sorts of being who are in some sense responsible by our very nature, or we are not responsible at all.

What I am suggesting here is no doubt highly mysterious: I seem to be ruling out virtually every feature necessary for freedom—the existence of prior grounds of choice, the indeterminate choosing among those grounds—as themselves undermining free will. And this is highly counterintuitive. In my defense, I note only that these points *are* problematic; that is exactly why there have been ongoing, increasingly complex efforts to deal with them. If none of those efforts have taken permanent ground, of course, this does not mean that the process of adding complexity is misguided. I want only to suggest that perhaps a different strategy altogether may be needed. I will hint at it briefly in the next chapter, and develop it in full in my final chapter.

3

Experience, Deliberation, and the Construction of the Will

One way of attempting to address the problem of free will is to appeal directly to features of consciousness or deliberation. One could say: of course I am free, all experience is in favor of it, as Dr Johnson is supposed to have said. Of course he also noted that all theory is against. One way to defend free will, then, is to stick to experience. In this chapter I will argue that this is a mistake: experience is *not* for free will. Neither is the nature of deliberation itself.

A. Searle

Searle attempts to find freedom primarily in the experience of a "gap" in our experience of choosing and acting. As to how the experiential defense of freedom is to be defended ontologically, he asks: "there is no doubt that the gap is psychologically real, but is it otherwise empirically real?" (2001 269) He rejects the appeal to quantum indeterminacy and admits that it is not clear how we could exercise conscious causation within nature; but he concludes that this problem is really the problem of consciousness, so anyone who rejects libertarian freedom will be unable to explain consciousness as well and that difficulty seems to give the libertarian a certain leg up. Searle's position, then, is that if we establish that experience is on the side of freedom, we have won the major part of the battle; reconciling experience with reality is a secondary task. The real argument is

in his theory of rationality, which Searle presents as an attack on what the calls the Classical Model of rationality. What he means by this is largely Humean rationality, but he takes Davidson—mistakenly, I think—as his main target. The main point of the Classical Model that Searle seeks to refute (at least for my purposes) is the claim that our actions are caused by a combination of desires and beliefs.

The centerpiece of Searle's account is the claim that freedom consists of a gap: "'The gap' is the general name that I have introduced for the phenomenon that we do *not* normally experience the stages of our deliberations and voluntary actions as having causally sufficient conditions or as setting causally sufficient conditions for the next stage." (2001 50) Specifically, the gap is the experience of indeterminacy between our reasons for a decision or action and that decision or action itself. Though Searle thinks there is really one gap, he argues that we can locate it in three different places, in each of which the self must play the crucial role of connecting conditions to consequences; nothing else can play that role ("what fills the gap? Nothing. Nothing fills the gap" (2001 17)[13]). First, there is a gap between the reasons for acting and the decision made on the basis of those reasons; the reasons by themselves cannot be causally sufficient for the decision. Second, there is a gap between the decision and the action. Third, there is—in any extended action—a gap between the initiation of the action and the carrying out of that action to completion; for example, I might start writing a dissertation but then give up (not bloody likely), and finishing the project requires conscious continuation of the action on my part. In each case, the gap implies that, whatever the preceding initial

[13] There is an implicit reference to Sartre here: what Searle means by the "nothing" is that there are no causally sufficient conditions: the "nothing" is a way of saying that only the self operates in this gap.

conditions, the region between those and the final outcome is the domain of freedom. I want to look specifically at two arguments in this connection.

First, Searle insists that "when one has several reasons for performing an action, or for choosing an action, one may act on only one of them; one may select which reason one acts on." (2001 65) Searle takes up as his example the case of voting for a particular candidate, where I have a number of reasons to do so. "I may not vote for the candidate for all of those reasons. I may vote for the candidate for one reason and not for any of the others. In such a case, I may know without observation that I voted for the candidate for one particular reason and not for any of the others." (2001 65) My conscious awareness of choosing a particular reason for action demonstrates that I am free with regard to my reasons—they do not cause how I act. Rather, I choose how I act on the basis of them. "If we think of the reasons I act on as the reasons that are *effective*, then it emerges that where free rational action is concerned, *all effective reasons are made effective by the agent*, insofar as he chooses which ones he will act on." (2001 66) Searle makes it clear, then, that he is replying to the concern I raised earlier in relation to Davidson, that is, the question of what accounts for the causality, or efficacy, of a reason. Searle's point is that a reason cannot, by itself, cause an action—an agent has to act *on* that reason.

Searle's goal with this argument is to show, contra Davidson, that reasons are not causes. In that sense, the argument clearly falls far short of its mark. It is based on a common misunderstanding of what Davidson means by his claim that reasons are causes.[14] The misunderstanding involves trying to show that reasons do not provide

[14] Goetz, for example, makes a similar point in claiming that "a reason provides a basis for the agent acting in one way as opposed to another (or not acting at all) without *causally* determining the agent to act (or not to act) in the way in which she does." (1988 312)

sufficient causes for action.[15] But of course Davidson never claims that reasons are sufficient causes for action. First, reasons appear as causes only retroactively. Before an action is committed, it cannot, on his account, be predicted on the basis of the agent's reasons. The causality of reasons functions as an explanatory, and not as a predictive element. Second, Davidson actually *stresses* that the reasons for which an agent acts are not sufficient causes. He points repeatedly to the various problems of irrationality and wayward causal chains that make it impossible to find the sufficient rational causes of an action. And he is very clear on this point: no given reason can be a sufficient cause for any action because the agent acts not in light of a single reason, but after consideration of many reasons. Of course, this might still imply that even though no particular reason is a sufficient cause of an action, the reasons making up an agent's total motivational set might serve as sufficient causes, and that would provide a challenge for freedom. But even that suggestion is fully rejected by Davidson, on the grounds that it would make irrationality (especially akrasia) incomprehensible. He points out that "knowing the intention with which someone acted does not allow us to reconstruct his actual reasoning." (Davidson 1980e 98) Or, in another place, "every judgment is made in the light of all the reasons in this sense, that it is made in the presence of, and is conditioned by, that totality." (1980d 41) The point is that an agent's act is based on a judgment, itself caused by some reason. But the question of *which* reason ends up causing the judgment, and for that matter which reasons out of the entire set of an agent's reasons are even

[15] See, e.g., Searle: "cases of actions for which the antecedent beliefs and desires really are causally sufficient, far from being models of rationality, are in fact bizarre and typically irrational cases" (12); or again: "I can tell you why I am doing what I am now doing, but in telling you why, I am not trying to give a causally sufficient explanation of my behavior, because if I were, the explanation would be hopelessly incomplete… because in specifying these causes, I do not give you what I take to be causally sufficient conditions." (Searle 2001 69)

considered, or taken to be important, by the agent is left completely open. The total set of an agent's reasons, in turn, conditions the judgment—insofar as it provides the setting in which deliberation must take place—but it cannot determine the outcome.

Not only does Davidson have a response to the arguments against him, but in fact his position is stronger than Searle's, because it has a way of explaining actions that does not terminate simply with "the agent did it." I mentioned earlier that, in discussing the question of the agent's role in making a reason effective, Searle is addressing a gap in Davidson's account. That much is true. But Davidson does not deny that the agent has a role in determining which reason causes his action—he rather insists on that fact. It is only that he does not explain how that happens. Searle attempts to fill the gap by, well, inserting a gap. But our alleged ability to choose which reason we actually act on does not really show that we are free in the strong sense Searle intends. For one thing, we frequently do *not* know which of our reasons we acted on; though Searle is certainly not the only one to think that we do, I cannot imagine the piece of phenomenology that would confirm that intuition. We often believe we are acting on one reason rather than another because we deceive ourselves, and frequently, when we think about it, we change our mind about which reason we *really* acted on. In an interesting turn, this is a much bigger challenge for Searle than for Davidson. The latter recognizes that we are frequently wrong about our own reasons, but this does not touch on the claim that reasons serve as causes: we are frequently wrong about the causes of all sorts of things (like weather patterns), but this does not generally lead us to think that those things are uncaused. (Davidson 1980a 18) But since Searle's argument is entirely about the *consciousness* of freedom, I do not see how it could deal with the challenge.

There are two upshots of this discussion: first, there is no reason to think that reasons are not causes; the claim that an agent chooses which reasons to make effective does not make the reasons any less causal, provided we understand that a reason is not a sufficient cause. Second, this strategy for demonstrating that we are free by taking a first-person perspective and peering into our own consciousness as opposed to some third-person standpoint ultimately fails to show anything about our freedom: we may think we are free and feel we are free (if there is such a feeling), but this can be illusive—the conclusion that we really are free, even psychologically, is only as strong as the certainty of our self-knowledge. I want to dwell on this second point in a slightly different context by discussing Searle's second major argument together with Korsgaard's approach.

Searle seems to think that our freedom is, really, an analytic truth given that we are rational: if we are capable of exercising rationality, this implies that the use of rationality must make some difference to how we act. But if rationality is to make a genuine difference, then it must be possible for us to act otherwise than we actually do. This is a roughly Kantian argument, and Searle works it out through a strategy frequently ascribed to Kant.[16] The important point, on Searle's account, is that we can deliberate only in light of the appearance of alternate possibilities: deliberation as such only makes sense on the assumption that the deliberation has a role to play in determining which way we end up deciding, and this presupposes that there must be more than one thing we could decide. The point, though, is not simply that there are different ways that I *could* decide (since a determinist can easily accept that but reject the further claim that what I

[16] "Kant pointed this out a long time ago: There is no way to think away your own freedom in the process of voluntary action because the process of deliberation itself can only proceed on the presupposition of freedom, on the presupposition that there is a gap between the causes in the form of your beliefs, desires, and other reasons, and the actual decision that you make." (Searle 2001 14)

decide is ultimately up to me), but that my deliberation is instrumental to the final outcome; that means that I must assume that I am free in order to be able to deliberate at all. Does this conclusion really follow? Searle suggests the following: "If I really thought that the beliefs and desires were sufficient to cause the action then I could just sit back and watch the action unfold in the same way as I do when I sit back and watch the action unfold on a movie screen. But I cannot do that when I am engaging in rational decision making and acting." (2001 71) Here is an even more graphic display of the point: "Suppose you go into a restaurant, and the waiter brings you the menu. You have a choice between, let's say, veal chops and spaghetti; you cannot say: 'Look, I am a determinist, che sarà, sarà. I will just wait and see what I order! I will wait to see what my beliefs and desires cause.' This refusal to exercise your freedom is itself only intelligible to you as an exercise of freedom." (2001 14)

This is a very old argument—not really an argument so much as a misleading intuition pump. But clearly it has nothing to do with the point at issue. A determinist can simply respond that we are determined to deliberate and the actual process of the deliberation is determined, so that any step from our deliberating to our being free cannot be a logical one. The convinced determinist need no more sit back and watch what he ends up doing than the convinced libertarian needs to spend every conscious moment of his life making decisions. Searle's point seems, at least on the surface, to depend on the idea that deliberation includes an experience of freedom as an internal element. This is not a new idea, but I confess that I have never been able to see its force. If "the experience of freedom" is just a redescription of "the experience of deliberation," then the claim is question begging. If, however, there is some separate and unambiguous

experience of freedom in this sense, I confess that I lack it. "What is it like to be indeterministically free?" I don't know, but it doesn't seem to be much like the experience of seeing red. Maybe it is a bit more like being a bat.

But let me try to reconstruct the argument more charitably. If I am faced with a choice, I do not know what I am going to choose until I choose it. Thus, I am at least uncertain about the outcome of my deliberation. This means that I do not experience myself as being driven toward a particular outcome. Furthermore, the arguments against this experienced indetermination come from outside—from assumptions about how the world works. But from a first-person perspective, there is at least no experience of determination. Searle does in fact seem to take an approach along these lines: "The question 'Why did you do it?' asks for a totally different sort of answer from the question 'Why did it happen?'," and this distinction leads to the advice to "always look at phenomena such as rational behavior and its explanation from the first-person point of view, because they have a first-person ontology. They only exist from the first-person point of view." (2001 85) This is an interesting point, and Searle raises it in reply to Nagel: in deliberating, we *cannot* see our actions as mere events in the world, because we see these actions from a different perspective—a first-person perspective—than the perspective from which we observe events.

Furthermore, Searle raises the interesting suggestion that the first-person perspective has an *ontology* that differs from the deterministic, third-person, perspective. Unfortunately, Searle doesn't quite develop the implications; in fact, he seems to cancel them out in claiming that "the gap might be an illusion," though he immediately adds that "it is not a belief we can give up." (2001 71) The problem is this: if the gap (or freedom)

has a different *ontology* from that of the sciences, if this is a first-person ontology, and if belief in freedom cannot be avoided from the first-person perspective, then what can it mean to say that it might be an illusion? I cannot see how Searle could avoid the problem except by suggesting either that freedom might be doubted from the first-person standpoint, or that there is a meta-ontology within which the first-person and third-person ontologies must ultimately be reconciled (where the third-person ontology is assumed to have the upper hand). The second approach would, of course, dull the force of the word "ontology" (and, in fact, make it completely useless), but the first would seem to completely undermine Searle's entire argument. Below, in conjunction with a similar critique of Korsgaard, I will argue that the first approach is actually the better strategy here.

B. Korsgaard

Although written several years before Searle's account, and providing a clear influence on his version, Korsgaard's argument is far more subtle. Korsgaard defines freedom in terms of the reflective capacities of our minds: to act, we must have a reason for acting. But having a reason is not a matter of just picking some desire or other and going with it. In fact, insofar as we are reflective, we cannot just do that. We have to decide whether or not the desire (broadly construed) provides a sufficient reason for action; that is, we must evaluate it and either endorse or reject it. "The reflective mind cannot settle for perception and desire, not just as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward." (1996b 93) Our

freedom, then, is the freedom from external determination of the will, and it is a freedom because our reflective nature makes it possible for us to question any desire that presents itself as a candidate for determining our action. Having questioned, we *can* reject the desire, so that nothing determines the will to act except its own rules. Korsgaard's argument is not that this proves that determinism is false, but rather that the truth or falsity of determinism has no bearing on our freedom.

Korsgaard suggests that determinism might seem to pose a problem for freedom by giving us the knowledge that, even if we thought we were free, in fact we could not have done otherwise. But that still would not challenge freedom; at most, Korsgaard argues, it seems like a challenge to responsibility. It is not a challenge to freedom because freedom does not operate within the same sphere as theoretical knowledge. "The freedom discovered in reflection is not a theoretical property which can also be seen by scientists considering the agent's deliberations third-personally and from outside. It is from within the deliberative perspective that we see our desires as providing suggestions which we may take or leave." (1996b 96) In other words, there are two perspectives or, as Korsgaard famously argues, "two standpoints, or ways we have of looking at things… they represent a practical and a theoretical viewpoint" (1996a 185): on the one hand, we can look at ourselves from a scientific, theoretical, third-person perspective. On this view, we are fully determined, and the point of this perspective is to explain why our actions occurred and to predict future actions. On the other hand, we can look at ourselves from the practical, first-person perspective of deliberation, and there we are trying to decide not why we act, but how to act. The concepts of freedom and determinism thus apply to two different perspectives or standpoints and cannot conflict with each other.

To make the point clearer, Korsgaard invites us to imagine that scientists have taken control of our brains as part of an experiment. In this situation, you know that the scientists are controlling your actions as well as the thoughts leading up to those actions. But while your knowledge that you are being controlled by scientists can influence your decision (you might try to do something unpredictable, hoping to outsmart the machine, or might try to stop making decisions altogether—in this earlier piece Korsgaard is still using that version of the argument), but "in order to *do* anything, you must simply ignore the fact that you are programmed, and decide what to do—just as if you were free. You will believe that your decision is a sham, but it makes no difference." (1996a 163) This is a subtle move; unlike Searle, Korsgaard openly rejects the idea that we must believe ourselves to be free. "The point is not that you must *believe* that you are free, but that you must choose *as if* you were free. It is important to see that this is quite consistent with believing yourself to be fully determined." (1996a 162) But I am not sure that Korsgaard quite succeeds in holding on to this move. Her point "is not about a theoretical assumption necessary to decision, but about a fundamental feature of the standpoint from which decisions are made. It follows from this feature that we must regard our decisions as springing ultimately from principles that we have chosen, and justifiable by those principles. We must regard ourselves as having free will." (1996a 163)

Korsgaard's argument shows that determinism does not pose a threat to freedom in the sense that it cannot—or should not—change how we act; at least, unless we react to the idea of determinism in a highly irrational way. But this argument does not seem so much to say anything about our freedom, but rather about our psychology, particularly

our psychological reactions to determinism.[17] Moreover, the appeal to the first-person as opposed to a third-person view of ourselves as determined seems to miss a major feature of Nagel's challenge: Nagel argued that, even though we cannot let go of freedom, the objective view of ourselves forces doubts on us. Taken far enough, it involves a radical skepticism about freedom. But the objective view, for Nagel, is not the third-person view: it is thoroughly first-personal. His point is not that we only doubt freedom when we look at ourselves from the outside, but that we can—and in the course of reflection must—come to see ourselves from the outside while remaining within the first-person perspective.[18] But, on a related note, Korsgaard seems to miss a deeper threat to freedom. The threat is not that we will be unable to deliberate rationally; the threat is that the tools of that deliberation, the values and reasons we already have, will serve as the grounds based on which we deliberate; but those values and reasons cannot themselves be freely chosen by us, at least not all the way down, because for that we would need to freely create ourselves from outside the world. I might act as if I am free, and I might try to do my best to have good reasons for my actions, but I can also recognize that it is impossible for me to fully subject *all* my relevant reasons to scrutiny. I may regard myself as free in the sense that knowing determinism to be true would not alter my behavior in any way, but this is neither equivalent to nor sufficient for regarding *all* the principles on which I act as chosen.

From Nagel's perspective, the threat to freedom is that my decisions, at bottom, cannot spring from something I have chosen or created. Both Nagel and Korsgaard seem to agree that, if we act on sufficiently objective reasons, we will have a kind of freedom.

[17] Similar versions of these arguments are raised in Chapter 3 of Guevara (2000).

[18] For a more developed argument along these lines, see Nelkin (2000).

But if our decisions are necessarily conditioned by features of our character, and these are furthermore not features we could have freely chosen, this undermines our ability to act on those principles that we might consider to be the best, or most objective, principles of action. Our reflective nature, certainly, is not enough to counter this threat: if all our actual reflection is conditioned by features we have not chosen, then whether or not we will act on principles that we *have* chosen—even if there are such—will be merely a matter of luck. This is a variant of the problem I earlier raised with regard to libertarian theories: if our character is not chosen, both the freedom and the responsibility of our actions are undermined.

Finally, what does it mean to say that I must deliberate "as if" I were free? If this claim says something meaningful, it ought to be possible to imagine a contrary alternative. That is, can I deliberate "as if" I were not free? For both Searle and Korsgaard, that alternative is incoherent. But, in fact, I think there *is* an alternative. Both Searle and Korsgaard present the argument that we cannot simultaneously, or from the same point of view, see ourselves as free *and* determined. And it is this version of the argument that falls to Nagel's analysis, because it seems that we can, from the first-person perspective, see our actions as merely events in the world. In considering the alternative to the "as if" scenario, we can achieve two things: first, we undercut the pretensions of these arguments to open a realm of genuine freedom within the first-person perspective by introducing a first-personal determinism. But second, we can modify the arguments in light of this objection to undercut Nagel's eliminativism. Here is how.

The argument for freedom depends on the idea that, from the first-person perspective, we cannot see ourselves as determined: we *must* see ourselves as free. But I don't think this is true, because consciousness is not transparent. Take the following scenario: I have applied to several graduate programs. Months go by with no reply, and I become anxious and pessimistic. Finally, a representative from program A calls to inform me that I have been accepted. As a result, I immediately develop a strong liking for this program. A week or so later I receive an acceptance letter from program B. Now I have to make a decision. I deliberate for weeks. I visit both campuses, compare various impressions, breadth of faculty interests, financial packages, study-abroad opportunities, requirements, placements, and so on. Based on these criteria, I finally decide in favor of program A. Now what I have just undergone is a genuine deliberation: I did consider various factors, and I chose which ones of them are effective in making up my mind. That is, I chose the reasons I act on and acted based on those reasons. But did I? Isn't it possible that the warm feelings initially engendered by the phone call from program A inclined me toward program A, coloring all my other considerations? In other words, the decision I made may well be biased. The libertarian could respond that the decision is still not *determined*. But how do I know that? In one way, I do feel that I genuinely deliberated and reached a decision based on my deliberation. But at the same time, I feel that perhaps I had already decided prior to the deliberation, prior to considering any of the factors, and that the deliberation process was not so much a process of deciding as a process of retrospectively rationalizing my decision to myself. Can I be sure that my actual decision was not really determined but merely biased? I do not see how.

This may be a somewhat extreme case, but such cases do occur[19]: cases in which we undergo deliberation, but at the same time entertain some doubt about whether the deliberation is the determining factor in the decision. In other words, it *is* possible to (1) go through a process of deliberation, and (2) doubt whether that deliberation is a decisive factor. I am not suggesting that such cases necessarily involve a loss of freedom. What they *do* involve, however, is a case where we do not act "as if" we are free, at least in the total sense Searle and Korsgaard seem to suggest. Though Searle does not have this option, Korsgaard might answer that even if we believe ourselves to be determined, we must still deliberate "as if" we are free: we have no choice but to go on deciding how to act, regardless of whether or not we think this deliberation will be effective. That, in fact, is precisely what Korsgaard says. But we must distinguish between the following sorts of deliberations: (1) searching for the best reasons on which to act, (2) searching for the best reasons to support a pre-decided course of action, and (3) searching for the best justification (perhaps in a moral vein) for a pre-decided course of action. And there seems to be a difference between searching for causally effective reasons on the one hand, and searching for rationalizations or justifications on the other. The difference is not merely one of how we see the very same activity of deliberating, but must change the form and course of that deliberation itself. The threat is that we might see ourselves as spectators rather than as agents.

Both Korsgaard and Searle could respond that cases such as I have described are rare, and that maybe we can isolate those situations in which we do suspect that our

[19] I suspect, in fact, that this is what the vast majority of our decisions are like—which may be one reason Kane and van Inwagen exclude most of our decisions from the domain of freedom—but I won't press the point. Merleau-Ponty seems to me to be suggesting something similar at the beginning of his chapter on "Freedom" in *The Phenomenology of Perception*.

decisions are determined prior to deliberation and take only others as paradigms of deliberation and therefore freedom. But this might be harder than it seems. What cases of the sort I indicate suggest is that it is possible to see ourselves as both deliberating *and* determined at the same time, and to see ourselves this way from the first-person perspective. If this is possible, there is no reason I can see, other than dogmatic assertion, to insist that the same sort of prior determination does not occur in the cases where we do not experience ourselves, prior to deliberation, as inclining toward one of the alternatives. Referring to this possibility as the second of three objections to his theory of the gap, Searle offers: "maybe the unconscious psychology overrides the conscious experience of freedom in every case. The psychological causes may be sufficient to determine all our actions, even if we are not conscious of these causes." And then he counters, in typical Searle style: "I have nothing to say about [this objection], because I do not take it seriously. There are indeed some cases where our actions are fixed by unconscious psychological causes—hypnoses cases for example—but it seems incredible that all our actions are like acting in a hypnotic trance." (2001 63-64)

I think Searle is simply missing the force of the argument, and this for several reasons. First, the case I have suggested is clearly not a case of hypnosis or even remotely similar. Second, the causes involved need not be unconscious. Consciousness is not completely transparent, and there may well be all sorts of thoughts, motives, and reasons that influence our actions, and that are conscious, but that are not explicitly conscious (rather pre-conscious, in Freud's sense). This latter point can be bolstered in the following way: sometimes we are quite sure that we are acting on a particular reason, but in retrospect change our minds about what our reason was. Should this suggest that the

motive we *now* think was dominant is a new invention? That it was not present at the time? Third, it is not necessary that *all* our actions be fixed by underlying motives and reasons in this way. If *some* of our actions are, then this is enough to at least create doubt about our other actions. My argument need not establish that we simply lack self-knowledge. It need only put the certainty of that knowledge into question. If we can sometimes be wrong about having genuinely open choices, then it is at least possible that we are always wrong. That is: we find no refuge from the threat of determinism within the first-person perspective. We need not, in order to deliberate, do so "as if" we were free. In fact, even in cases where I do recognize a very strong prior inclination toward one of the alternatives, I might still feel—because of my reflective nature, and because of a need to justify the final decision, even if only to myself—a need to weigh my options, to deliberate, despite the strong suspicion that the outcome is determined. Of course I could, just to spite the determinist, choose to act against my prior inclination; but that would make my action utterly irrational.

Nagel's argument that free will is actively threatened from within the first-person perspective thus finds support in this view. As I have been arguing, I do not think Searle and Korsgaard succeed in demonstrating what they want to: that we must act as if we are free, or with the belief in our freedom, so that at least at the level of psychology we are free from determinism. Having pointed out the role of reflection in our conception of freedom from a *deliberative* rather than theoretical perspective, Korsgaard continues: "You will say that this means that our freedom is not 'real' only if you have defined the 'real' as what can be identified by scientists looking at things third-personally and from outside." (1996b 96) Especially in some circles, this view—that determinism cannot be a

threat to freedom except from an impersonal perspective—has a certain amount of authority; I think it is mostly bunk. And it falls to the notion of the "blind spot" pointed out by Nagel. As mentioned earlier, Nagel thinks that the objective view at first gives us hope for autonomy: by widening the scope of our self-knowledge, it also widens the scope of the motives we can subject to scrutiny. "But this objective self-surveillance will inevitably be incomplete, since some knower must remain behind the lens if anything is to be known." (1986 127) Thus, though we may strive to attain complete self-knowledge, it will always be beyond our grasp. Our view of ourselves is "essentially incomplete": "The incomplete view of ourselves in the world includes a large blind spot… that hides something we cannot take into account in acting, because it is what acts." (1986 127) This blind spot drives another nail into the coffin of autonomy. Insofar as the idea of autonomy involves the idea of being able to know *all* the motives that influence our decisions and actions and subject those motives to reflective scrutiny, the blind spot presents an insurmountable problem: if we cannot have complete self-knowledge, then "our actions may be constrained by an influence we know nothing about. This might be either something we could successfully resist if we did know about it, or something we wouldn't be able to resist even then, but which we also couldn't accept as a legitimate ground for action." (1986 128)

The final step in the argument against autonomy along this line is that "we can't decisively and irrevocably endorse our actions, any more than we can endorse our beliefs, from the most objective standpoint we can take toward ourselves, since what we see from that standpoint is the incomplete view." (1986 128) This, in essence, is the argument I have raised against Searle and Korsgaard—but also against the libertarian tradition as

such—as the real and overlooked threat posed by determinism. Moreover, my attempt was to expand this threat from an epistemological necessity into a real psychological possibility. But we can also reverse the argument against Nagel: what acts is what we cannot take into account in deliberation. There is therefore the possibility—it is *only* a possibility at this point—of responding that the objective view cannot decisively exclude agency and freedom, because it cannot get at the blind spot. There is—as I noted at the very beginning of my discussion of Nagel—a limitation to the objective view. That limitation comes from the fact that the objective view is always taken *by* someone; it is not, as Nagel admits, a view from nowhere. As transcendental traditions insist, the limit on any objective view is the result of the fact that it is a view, that is, that it is always from some perspective. Whether the source of that perspective can be a source of freedom is a point I will put off. While Searle and Korsgaard cannot use appeals either to our consciousness of action or the nature of deliberation to show that freedom is guaranteed by the first-person standpoint, the standpoint itself might provide such an opening.

C. The Will

A final point of disagreement between Korsgaard and Searle concerns the nature of the will. A recurrent objection to Korsgaard's argument in *The Sources of Normativity* is something like the following: Korsgaard claims that to act on a reason is necessarily to act on a self-given law. But, the objection goes, it is possible for human beings to act capriciously. I can decide to act on a whim without committing myself to any law that

requires me to act on all of my whims, or even on this whim every time it occurs. This is a common tendency for us, and to deny it seems a bit silly. Korsgaard's reply to this line of criticism is ingenious, but as with the argument about freedom, it seems to vacillate between describing transcendental conditions of acting and some apparently contingent points about human psychology, and it appears to move a bit too freely from the former to the latter. Nevertheless, there is a core to her argument that I hope I can interpret correctly in order to extract its truth. Here is where a number of the strands of this chapter will, hopefully, come together.

Korsgaard's argument appears to be something like the following: We can understand willing by analogy with the Humean notion of a cause. A cause involves a constant—i.e., regular—conjunction of events. If we did not have law-like regularity, we would be unable to identify something as a cause at all; we would not be able to distinguish two events following each other from two causally related events. The will must work in an analogous way. "Willing is *self-conscious* causality, causality that operates in the light of reflection. To will is not just to be a cause, or even to allow an impulse in me to operate as a cause, but, so to speak, to consciously pick up the reins, and make *myself* the cause of what I do." (Korsgaard 1996b 227) But this means that I must be able to distinguish between myself causing an action and one of my desires causing my action through me. "I am not the mere location of a causally effective desire but rather am the *agent* who acts *on* the desire. It is because of this that if I endorse acting a certain way now, I must at the same time endorse acting the same way on every relevantly similar occasion." (1996b 228) This is clearly confusing; the two sentences do not seem to belong together in the order and relation that Korsgaard gives them. The first

claim seems to be a definition, or a metaphysical description of a self: a self is not a site of active desires, but is itself active with regard to those desires. But *if* a self is by definition active with regard to its desires, it becomes difficult to see why this requires me to endorse my actions now in any more general form. That is, a move from "is" to "ought" is implied here with no clear explanation, and I think this problem persists through Korsgaard's early account. On the one hand, if I do not will generally, then I am not an active self. On the other hand, my being an active self requires me to will generally. Something here is in the wrong order, and I am not convinced that it can be fixed in the way Korsgaard intends.

Searle, in fact, is convinced that it cannot be. He points out that Korsgaard is indiscriminately mixing epistemic conditions for identifying a cause with ontological conditions of being a cause. As she tells us:

> Just as the special relation between cause and effect, the necessitation that makes their relation different from mere temporal sequence, cannot be established in the absence of law or regularity, so the special relation between agent and action, the necessitation that makes that relation different from an event's merely taking place in the agent's body, cannot be established in the absence of at least a claim to law or universality. So I need to will universally in order to see my action as something which *I do*. (1996b 228)

This is more than a little odd. As Searle points out, regularity is an epistemic condition of our being able to recognize a relation of causality; but the lack of regularity does not guarantee a lack of causality. The fact that we do not see a relation does not mean that the relation is not there.[20] Thus, Searle rejoins that "we can say that from the third-person point of view it is indeed an epistemic requirement on my *recognizing* somebody's decisions as truly his considered decisions, as opposed to his capricious and whimsical

[20] It is only a little ironic that Searle is here using precisely the sort of argument a Davidsonian would use against him.

behavior, that they have some sort of order and regularity. But it does not follow that in order to *be* his decisions, they have to proceed from a universal law that he makes for himself." (2001 155) Searle has a genuine complaint here, but he is clearly missing some important words in Korsgaard's argument: "I" and "my."

Searle's criticism would be completely right if Korsgaard's argument were that we could not identify the actions of others as *their* actions in the absence of regularity. But this is not her argument. Her argument, rather, is that if we did not adopt generality—or normative regularity—into *our* reasons, then *we* could not identify *ourselves* as agents. Thus, Searle's references to the third-person point of view are out of place: Korsgaard's point is explicitly about the first-person perspective. I suspect that Searle's mistake is not accidental: he intentionally misconstrues Korsgaard's position because he wants to reject the will: he wants to argue, instead, that regardless of whether we act on principle or on whim, "the experience of the gap can be the same in both cases." (Searle 2001 156) As I have already argued, the whole matter of the "experience of the gap" is a mysterious deal—in my view it is a theoretical misinterpretation of our actual experience—and cannot be used to support any argument. Furthermore, Searle has another stake in this debate: he wants to argue against Korsgaard's notion that the self somehow makes or creates itself by willing universally. If the self does so, "this is a totally different notion of the self from the one I am now expounding. [She] must mean we create our character and personality. The point I am making now is not that action creates a self, but that action *presupposes* a self." (2001 87) Searle's presupposed self is actually a completely shallow formality; in fact, he says nothing about it other than that it somehow chooses which of its reasons for action will become effective; it is, effectively, a Cartesian self that

exercises pure willing in a vacuum, with no consequences for its identity. If that notion of a self is not incoherent—and it isn't—it is at least incredibly uninformative. Searle goes so far as to simply attribute freedom to this self by definition, because a free self is needed to account for the experience of the gap. But this is just piling suspicious arguments on top of each other. We can, however, get the kernel of a legitimate complaint from this argument: it is not clear how the self can create itself without already being a self. I will come back to this.

Korsgaard's argument, despite its obscurity, is far more interesting than Searle allows. An agent, for her, is not an abstract entity that somehow works on reasons, but rather an entity that, by definition, is self-creating. Searle dismisses the idea because, clearly (for him, at least), creating our "character and personality" is a merely contingent matter, which comes only after the *real* issues of free will have been settled. But that is a mistake, and it is in this confrontation that we see Korsgaard bringing together the strands of compatibilism and libertarianism to transcend the typical limitations of both. I want to bring out, at least in general form, how she accomplishes this, and I want finally to relate this move to the notion of temporality at which I have been hinting throughout.

It is true, despite Searle's various errors, that Korsgaard does seem to conflate epistemic conditions for identifying agency with ontological conditions of being an agent. She does not confuse these in the way that Searle thinks, however, although this too is confusing, because Korsgaard moves quite quickly from an account of why regularity is needed to identify causality to an account of why we need universal principles to underlie agency. But this is not where the odd part of her argument is to be found. Instead, what's odd is the ease with which Korsgaard moves from what is required for us to be able to see

ourselves as agents, to what is required of us *as* agents. Let me quote a few typical remarks.

First, there are those places where Korsgaard seems to suggest that universal principles are needed for us to be able to identify ourselves as agents:

> Nagel misses the point when he says that regularity does nothing to establish the causality of my will. What it does is establish my own ability to see myself as having a will, as having the kind of *self-conscious* causality that *is* a rational will. (1996b 229)

> I cannot regard myself as an active self, as *willing* an end, unless *what I will* is to pursue my end in spite of temptation. (1996b 231)

But then Korsgaard seems to also claim that generality is needed on an ontological level for the agent to *be* an agent:

> The function of the normative principles of the will, in particular, is to bring integrity and therefore unity—and therefore, really, existence—to the acting self. (1996b 229)

> If I change my mind and my will every time I have a new impulse, then I don't really have an active mind or a will at all. (1996b 232)

Perhaps what Korsgaard means is that one cannot be a self without being able to identify oneself *as* a self; she seems, in fact, to suggest something of the sort: "we impose the form of universal volitional principle on our decisions in our attempts to unify ourselves into agents or characters who persist through time." (1996b 229) But I do not think this really works, because Korsgaard wants to make the act of imposing general principles on oneself into a conscious, self-aware act of self-creation. This is the notion that Searle rebels against, and rightly so. Phrased without the baggage of the "gap," Searle's argument is essentially like this: my experience of my self is not altered by whether I act on principle or on whim. A capricious action is still identifiable as an action; if I have a tendency to act capriciously, perhaps I will see myself as lacking consistency and others

may see me as unreliable, but my sense of self will not be shattered. That seems right. The difficulty is that Korsgaard postulates essentially two selves: an "ephemeral self" that attempts, through making and following general principles, to create an "active self." Despite noting the character of paradox here, Korsgaard does not do nearly enough to resolve it; in fact she exacerbates it, because the phrasing of her remarks about being able to recognize ourselves as selves, or as agents, implies an empirical psychology: it is as if an ephemeral self, a self faced with decisions, feels a psychological need to make general principles—to create a will—in order to see itself as an agent, i.e., as in charge of its desires rather than subservient to them. At the same time, Korsgaard seems to recognize the oddness of this claim, and she seems to reject it, insisting that the real problem is "whether the *active* self can coherently be conceived as ephemeral." (1996b 230)

I hope the outlines of Korsgaard's account have become apparent in my discussion of the difficulties. Dwelling on difficulties first is poor expository strategy, but I do not believe that the account itself is coherent; there is thus no way of summarizing what strikes me as right about it without first acknowledging that my version is not faithful to the original and why. The difficulty is that if Korsgaard's claim is taken as a matter of empirical psychology then the account will not work. To work, it has to assume the ontological position: a self is an agent, or becomes an agent, by making universality into a feature of its decision-making. By having—or making for itself—a will, the self separates itself from its desires and becomes active with regard to them. This is not quite Korsgaard's account because this description of the self is not in terms of its desire or need to identify itself as an agent; it is about the need of a self to *be* an agent; but this is not a point of psychology or, if it is, then it is a matter of transcendental psychology: a

self must, to be a self, make its decisions into its will. I am saying that this is a matter of transcendental psychology because acquiring a will—or making a will for itself—is a condition of possibility of the self's becoming an agent. And it is on these grounds that I cannot accept Korsgaard's account as given. She seems to present the choosing of universal principles as itself an act of agency, but that gives rise to the infinite regress that the entire tradition, compatibilist and libertarian, has tried so hard to avoid: we must have a non-agential account of acquiring agency, or we will never have an account of agency at all. I am not here giving a solution to the problem; I will attempt to do that in my last chapter. Here I merely note the need for an account of agency that is not itself agency-dependent.

D. A First Shot at a Solution

I want to point to the three most important implications, as I see it, suggested by Korsgaard's account.

1. Korsgaard's account incorporates the insights of compatibilism, despite her explicit rejection of Davidson: she repeatedly claims that the point is to separate the self as agent from the self as a location for effective desires, and she makes two negative remarks about "anomalous" causes and desires. But the distance from Davidson is only superficial. It comes in two forms. First, the insistence that our actions are not caused by our desires. Second, the claim that the cause of our actions lies in the active self. Korsgaard gets at the issue by embracing Frankfurt's account, though with two modifications. First, the will is redefined: it is no longer taken to be identical to the

agent's effective first-order desire but, instead, is associated with the agent's evaluative and universal second-order principles.[21] Second, she adds the level of freedom Frankfurt's account was missing. The agent must still endorse her desires in order to be free, but this endorsement need not itself be necessitated by external causes—it can instead depend on universal principles (although, as I have argued, this point is problematic because Korsgaard takes the threat of determinism too lightly). And thus we come back to Davidson. Though Korsgaard argues that it is the self, not its reasons, that has the causal power, the distinction becomes far less dire when we recognize that the self—in its constitution as agent—is identified with its will. The causality, then, is not in the first-order desire, but in second order principles, or the reasons that an agent makes into reasons. This strikes me as a largely Davidsonian account, though one transformed via Frankfurt.

2. This appropriation of compatibilism allows Korsgaard to overcome many of the limitations of incompatibilist theories. The self's freedom and responsibility no longer need to be seen as tied to particular acts: they are tied, instead, to the universal principles taken up in those acts. That is, what is free is the agent's will; actions are free only in a derivative sense, insofar as they are the actions of an agent who is in turn already a being constituted by the possession of a will. As I argued earlier, libertarian theories suffer precisely from the fact that they take up a compatibilist concept of personhood and then append freedom and responsibility to that concept. Korsgaard gets around the problem by building freedom *into* the concept of an agent, but building it in as a freedom that goes all the way down, so to speak, so that it is not a matter of luck. Problems remain, which is

[21] This is a deviation from Frankfurt's early work. The relation between Korsgaard's (later) work and Frankfurt's later work will is discussed in Chapter 6.

why I will have more to say on the subject, but this is a major step. A resolution of the libertarian problem is implied in Korsgaard's account, and Searle is wrong to dismiss it so lightly. Searle's criticism—again, disregarding the problematic topic of the "gap"—is that we are equally free regardless of whether or not we act on general principles or on our whims. This is right. But Korsgaard's argument is not that we cease to be free (or to have wills, or to be agents) when we act capriciously, though her merging of epistemological and ontological conditions seems to imply this. We are, of course, free (if we are) regardless of how we act. But the point is that we could not act at all in any genuine sense—we could not have agency—unless we already accepted universal principles; that is, unless we already had a will. Or, to generalize the conclusion: the point is that the will is ontologically prior to the actions that issue from it. In a different register, we might say that Korsgaard combines the truth of compatibilism (that freedom and responsibility are internal aspects of agency) with the truth of libertarianism (that freedom and responsibility are ontologically irreducible to contingent features of empirical psychology, social norms, non-agential events, and so on). What allows for this combination is a reevaluation of agency against a background of temporality.

3. Temporality enters explicitly into Korsgaard's account. Of the theories we have examined so far, the temporality for the first time is a deep temporality. I believe Korsgaard gets the account backwards but, at the same time, she shows exactly why a theory of free agency requires a deep temporality, a point I have largely avoided discussing explicitly up to now. The account enters in Korsgaard's discussion of the universal principles that the self must adopt in order to be an agent. In explaining why a self must unify itself into an agent by taking up universal principles, Korsgaard states that

"the view of itself as active *now* essentially involves a projection of itself into other possible occasions." (1996b 230) In endorsing a desire to act as a reason, that is, as a universal ground for action, the self essentially creates for itself a temporally persisting identity. By deciding now that a certain desire is a reason, I decide simultaneously that this reason is universally valid: it is a principle that should be as binding on me in the future, and should have been as binding on me in the past, as it is now. An obvious criticism at this point would go like this: perhaps our principles are binding on us in the future, but what could it mean to say that they are binding in the past? Here is one answer: It is possible for me to act in a way I think is right but, in retrospect, to feel guilty about having acted that way. The reason is that in my past action I violated a principle to which I now hold. Although back then I was, in a sense, a different self with different principles, from my *current* vantage point I recognize that I was the same self and was thus subject to the same principles with which I now identify.

The argument that principles must be universal can be broken up into two parts. First, as we have already seen, a unity across time is needed in order to establish the self as an agent. Korsgaard draws out this idea by pointing to the role, first, of hypothetical imperatives and, second, our principles in general. A hypothetical imperative implies that, if we will a certain end, we must also will the means to it. Korsgaard's argument is that if we do not will the means—or, more specifically, if we give up the means every time that they seem too difficult—then we never really will any end. If this were to occur, we would not have agency, because we would be drawn each time only by the desire of the moment, which tempts us from our goal. The argument is not entirely sound: it seems perfectly conceivable that we might will ends without willing the means if they are too

difficult. This reflects that we do not will the ends very strongly, which is perfectly compatible with agency, but it does not, I think, involve a loss of agency. Yet Korsgaard's account has a deeper level, and this is another case where her language seems misleading. The point is that I cannot have a will at all unless I do, occasionally, commit myself to some ends. A self that genuinely cannot pursue goals is, clearly, a self that lacks agency. And so the better way of putting Korsgaard's point, I think, is not that following hypothetical imperatives is necessary for agency, but that the *ability* to follow hypothetical imperatives is. Something similar is true of principles: if, every time we are confronted with a difficulty, we deviate from our principles, then we have no principles at all. But the deeper point is that it is the ability or capacity of sticking to our principles, of resisting temptation, that is essential to agency. An agent, then, *is a self that is capable of unifying itself in time through its decisions in a given now*.[22]

Second, temporality is built into the very idea of being able to do otherwise. This is why I can act capriciously—I can violate my principles—and still remain free, even though it is the ability to have principles that constitutes my agency. "When we act self-consciously, we act under the idea of freedom: we think that we could act otherwise on *this* occasion. But that means that 'this occasion' itself must be conceived in general terms: it cannot be an ineluctable particular. You cannot say of an ineluctable particular that *it* could be otherwise." (Korsgaard 1996b 231) When I act reflectively, when I take my desire as a reason for action and so adopt a principle on which I act, I am not making a principle for any particular moment. That would make no sense. I am making a

[22] I am stressing this point because it is the substantive conclusion I want to draw from Korsgaard's account. As I am about to argue, however, I think it gets things exactly backwards: the will is not dependent on its choices in the now, but the reverse. Thus, I offer this substantive conclusion as a contrast to my argument in Chapter 6.

principle that applies to moments in the past and (more significantly) future that are in relevant ways similar to the present moment. The reason I must do this, as the argument just quoted shows, is that otherwise it would not make sense for me to say that I *could* have acted otherwise. If the reason I act on is only valid in this single instant, then it does not make sense to say that I can, or could, act otherwise, since I do not in fact act otherwise than I do. To say that I could act otherwise already implies that I am making a rule that applies not in a particular moment, but in a certain kind of situation; this situation, in turn, is only contingently tied to a given moment in time. That I can deviate from this rule at other times—in fact, that I could make a rule now for the moment at hand and immediately violate it—is possible. In fact it is implied in the universality of the rule. And this is Korsgaard's point: it is only because I have a rule, or because I choose my reason not for the instant but for the universal case, that it even makes sense for me to self-consciously act otherwise.[23]

How does this step introduce deep temporality into the account of freedom and why does it show the importance of that temporality? Korsgaard's argument, essentially, is that in making a reflective choice, I am simultaneously choosing a general condition, an attitude, that applies not just to the moment at which the choice is made, or to the moment at which the choice is carried out, but to any relevantly similar moment. The choice is an event that occurs at a particular moment in time. But the attitude that is chosen along with it is not a similar event. In fact, it is not an event at all. An event, as Davidson tells us, is a dated occurrence. But the attitude is neither dated nor an occurrence: it is a principle. More loosely, it is a disposition to act in a certain way on

[23] Of course I might act otherwise simply because I do not reflect. But since completely unreflective action is not taken by Korsgaard or by most free will theorists to be a free one, we can leave the issue of such actions aside.

similar occasions. On those occasions, if they arise, I may or may not act on the basis of this attitude, and my conforming or failing to conform to the principle will, each time, be an event. But the principle itself is clearly not an event; there are moments at which it is instantiated or not instantiated, but there is no moment or set of moments, not even a set of possible moments, to which the rule is confined. This is the notion of deep temporality I have been using: the relation between the choice and its relevant underlying attitude or motive or principle is precisely the relation between an event in time and a non-event, a constituent of the will, that is not in time.

Now, the second question: why is deep temporality important? As I have tried to show, Korsgaard's account opens the way to a functional theory of freedom that can also allow for a reciprocal notion of responsibility. Compatibilist accounts, I have argued, do not succeed because they lack genuine freedom in their attitudes; or at least in whatever attitudes ultimately underlie our choices. While the choices may be free relative to the attitudes, the origin of the attitudes seems to negate that freedom. I have tried to make this point in relation to both Davidson and Frankfurt. Libertarian theories, on the other hand, focus on indeterminacy, but since they accept the compatibilist account of the self—an account on which the motives and reasons among which the agent chooses are simply given to the agent from outside—they also seem to deprive the agent of genuine agency. Both the compatibilist and the libertarian, in other words, run into problems because their accounts are ultimately reducible to a shallow temporality. They take both terms in the choice/will distinction as events; even if the choice is free relative to the will, and even if the choice is genuinely undetermined in some way, its underlying will is still an event or product of events. This scenario leads directly to Nagel's eliminativism: if we

see both our choices and the relevant motivating factors behind them as mere events, freedom—at least in the sense of self-control or autonomy—becomes something incoherent. The deep temporality of Korsgaard's account opens the way out of the predicament: if our choices are freely chosen events, and if the attitudes behind them are not events at all and are also chosen, then Nagel's reduction of freedom to incoherence may be avoided.

But I have also argued that Korsgaard's account is problematic, because she fails to recognize the genuine threat posed by determinism and eliminativism. The threat, as I mentioned, is that the "blind spot," our lack of complete self-knowledge, might prevent us from being able to endorse our actions in a purely unbiased way. We might think that we are acting on the strongest reasons in light of self-chosen universal laws, but we might be wrong.[24] Furthermore, we might simply lack access to the best reasons for action. This is tied to another issue: Korsgaard does not establish the will as a guidance mechanism for actions, but the reverse. She is right to set up a deep temporality, but she accomplishes it in the wrong way. For Korsgaard we create our will by making choices that involve not simply the selection of a particular action, but a universal rule. But this, in turn, is similar to the libertarian model we have seen in Kane, on which a free choice involves both a choice of action and the assignment of relative weights to the reasons influencing that action. Korsgaard's addition of a temporal dimension on which, through its will, the self projects itself into other past and future circumstances, is significant, but it does not alter the model in a fundamental enough way. The will is chosen together with

[24] Since, as I have argued, freedom is needed to give us the ability to act on moral laws, I want to briefly raise the real Kantian threat to Korsgaard's account: the threat is that, although we might act legally, we would not be able to act *morally*. In a more common idiom: we might end up acting on the right reasons, but not *for* the right reasons. This threat is why freedom, for Kant, is not simply untouched by determinism, but requires the rejection of determinism, at least outside the phenomenal world.

the choice; while it has a different temporality, it does not establish a genuinely temporally unifying will unless the self decides to follow the general rules contained in this will in future circumstances. Korsgaard does attempt to deal with that difficulty by insisting that, if the self fails to maintain its will in other similar circumstances, then it ceases to be a self. But, again, this move is clearly insufficient, for Korsgaard strongly appears to be making a claim about empirical psychology and probably a false one at that.[25] She does at one point suggest that she is offering a transcendental argument, but I cannot see any particularly convincing evidence of that.

Despite the deep temporality suggested in the idea that a self commits itself to a will in its reflective choices, this alone is insufficient for either genuine freedom or responsibility. Korsgaard's step illustrates the role deep temporality can play in agency; but something further is needed, namely, a will that is *not* simply chosen within its actions, but one that underlies the actions. Unless we develop an account along those lines, I do not think we have a convincing account of freedom, and not one sufficient for either morality or responsibility. The will, or the attitudes constituting it, cannot be chosen within a temporal choice, because that merely reduces the will of the agent back to the particular choices made; in other words, the will vanishes into its choices instead of grounding them. This account, on which we choose our long-term dispositions by making decisions in the present, may seem satisfactory from the standpoint of surface psychological phenomena, but it is insufficient for any transcendental account that could ground responsibility. Its consequence is that the will fails to affect its choices in any

[25] As I mentioned earlier, it has to be false because one must already have agency in order to be able to actually commit oneself to universal rules. Korsgaard seems to reverse the order, but that reversal is untenable. This is separate from the point that one must be *able* to commit oneself to universal rules in order to be an agent at all.

way; it is rather the choices that determine the will. But if so, then the introduction of the will in this context fails to radically transform the standard libertarian account of freedom. What we need to resolve this problem is a way of having a free will without making it dependent on an event of any sort, whether a choice or not.

E. Time and Ownership

To briefly review the argument of the preceding sections: on the compatibilist position, freedom and responsibility require as a condition only that agents' actions follow from some underlying character or will. But if the will does not originate in some sense form the agent, what we are left with is a collection of states or events—willings, desiring, choosing, actings—connected to each other in some way. In order to insert the agent into this picture, one must make the agent into a site for the occurrence of events. There are various happenings, linked together through causal chains, so that some region of these events, entirely continuous with events outside the region, can be circumscribed and referred to as "agent." Here the agent is not really active at all, but only a recipient and medium for the happening of events. What acts is the world. The agent participates. To give the agent an *active* role, one must in some sense sever the links between whatever lies inside the region and what is outside of it.

This is the point at which libertarians come in. Typical libertarian accounts defend freedom of action, or at least the freedom of the effective deliberative process, and thereby seem to fall to the Humean objection: it makes no sense to hold someone responsible for an action that is in no way continuous with his character. The action is

then random; it is not an action at all, but a mere event. One libertarian attempt to deal with this problem is to state that the agent's character or will (including reasons, motives, and so on) is already in place prior to deliberation. The agent must decide on the basis of this character, but the decision is not determined by the character. Thus, whichever way the agent decides, his decision will be continuous with his character but, at the same time, free. But this still doesn't escape the eliminativist threat: if whatever choice the agent makes is continuous with his character, but his character is outside the domain of his freedom, then the freedom so attained does not seem to allow for the sort of radical responsibility that the libertarian wants.

I am setting conditions for freedom and responsibility that appear impossible to meet: on the one hand, agents must somehow choose on the basis of their will. On the other hand, their will must be unconditioned, that is, it must not be composed of elements that pre-exist the agent's choice. We are born into social structures that provide us with norms we adopt and we have various biological and psychological tendencies; if these wholly constitute our wills, then any choice we make will not be up to us. Perhaps we can choose among those tendencies of our wills, but we will be choosing on the basis of the wills. And if the wills are—to start with—not up to us, then choosing on the basis of them will not make them so. This combinations of requirements, however, seems to be incoherent: if the agent does not choose his will on the basis of pre-existing reasons, then his choice will be completely arbitrary; if he does, it is a product of impersonal events. And that is Nagel's point: what we are looking for in free will *is* incoherent. Responding to eliminativism, then, will require a further strategy. I have been suggesting that this strategy will have two features: ownership and deep temporality.

In the previous section I argued that we need an account of the will that (1) like Korsgaard's ranges over all times, or all choices that the agent might take up, and (2) unlike Korsgaard's is not reducible to individual events of choice but instead precedes and affects those choices. The former requirement is needed because otherwise we will not be able to get any genuine freedom out of the equation: we will be left with a sequence of events, none of which can properly be attributed to an agent. We need the latter requirement, on the other hand, because without it the temporal depth established threatens to disappear: if the agent creates his will at the same time as he makes a choice, then the will is threatened by dissolution from the outset. This is, in fact, already implicit in Korsgaard's account: the will must *actually* be effective with regard to the various choices of the agent; if it is constituted together with the choice, then it is only provisionally effective with regard to any other choices.

I have argued that the trick is to fit the agent into what is a mere sequence of events. How does one do this? Postulating an indeterminate choice does not seem to help. Either the choice is pre-conditioned, or it is entirely arbitrary. Neither option is helpful. But we can take a page from Frankfurt here: by identifying with their wills, agents can make those wills their *own*. And agents can *originate* actions only if they take those actions on the basis of their *own* will. Frankfurt's account of this will not do, however, since he makes ownership entirely a matter of luck. And that seems like a consequence of the view that ownership is a matter of a match between different states, each of which can be conceived as an event at some particular moment in time. So to develop a different view of ownership, we have to develop a different view of temporality, which allows our wills to be independent of particular events in time. The temporal independence of the

will allows it to affect choices in time without being reducible to them: it stops the regress involved in events that pre-exist and condition the will. At the same time, the agent's ownership of his will stops the regress involved in discovering the will to be reducible to non-agential conditions. Only by combining deep temporality and ownership do we get agents in a strong sense: entities capable of being the sole originators of their actions. To develop this account, I am going to start over. The attempt to work out an account of free will has hit an aporia. Instead of pursuing it further, I will now turn to the strongest theory of moral responsibility that does not presuppose free will.

4

Responsibility

A. The Impossibility of Moral Responsibility

Let me begin with Galen Strawson's "Basic Argument," which like Nagel's eliminativism is intended to show that moral responsibility is impossible, regardless of whether or not determinism is true. The simplest version of the argument goes like this:

(1) Nothing can be *causa sui*—nothing can be the cause of itself.
(2) In order to be truly morally responsible for one's actions one would have to be *causa sui*, at least in certain crucial mental respects.
(3) Therefore nothing can be truly morally responsible. (Strawson 1994 5)

The way the argument is supposed to work is quite simple. Strawson focuses on actions that are done for a reason (though presumably the same argument would be even stronger in the case of actions that are not), and states as a premise that in acting for a reason, "what one does is a function of how one is, mentally speaking." (Strawson 1994 6) But if this is true, it follows that, in order to be responsible for one's action, one must also be responsible for how one is, mentally speaking (Strawson later refers to "how one is" in this sense as "one's character or personality or motivational structure—one's CPM, for short" (Strawson 1994 9)). There are of course important ways in which we can shape how we are: we can evaluate and attempt to change our attitudes on the basis of our evaluations, and we can undertake to acquire habits that will change our CPM. The problem, however, is that in order to make such choices in the first place, one must

already possess "some principles of choice, 'P1'—preferences, values, pro-attitudes, ideals—in the light of which one chooses how to be." (Strawson 1994 6) In order to be responsible for the CPM we choose on the basis of P1, however, we must also have chosen the P2 on the basis of which we can choose P1. We have clearly arrived at an infinite regress.

What gives rise to the regress is the claim that, in order to be responsible for any action or state of character, we must also be responsible for whatever it is in our CPM on the basis of which we can choose that action or state of character. We could avoid the regress only if we could, at some point in the chain, be *causa sui*: if we could choose our CPM, and the resultant P_n, on the basis of no underlying CPM and P_{n-1}. But human beings cannot be *causa sui* in this way. Strawson brings the point home by rejecting two replies to the Basic Argument. One suggestion is that one's self—the self we hold responsible—is somehow independent of one's CPM. We can make choices on the basis of our CPM, but the CPM does not determine which choice we make; that part is up to our self. The response, however, is that this self, in order to be able to choose among alternatives, must still have some preferences, some P_n, on the basis of which to choose. Simply adding another level, a self, to the CPM model does not dissolve the problem; it merely pushes it back to that level.

Another argument, favored by libertarians like Kane, urges that indeterminism is the answer: what one does, on this view, may well be a "function" of how one is, but the choice between different alternatives, each of which finds some reason or motivation among the agent's existing CPM, is not determined by that CPM. Strawson's response here is to point out that indeterminism does not help. Either, once again, the choice is

made by the agent in some way (in which case it must depend on how the agent is), *or* the choice is completely random, and "it is absurd to suppose that indeterministic or random factors, for which one is ex hypothesi in no way responsible, can in themselves contribute to having any [responsibility] for how one is" (Strawson 2000 151). Either how one chooses is up to one, in which case it must depend on one's CPM, or it is random, in which case the result is a matter of luck, for which the agent cannot be responsible.

There are, no doubt, serious difficulties with Strawson's argument, particularly since he replies to opposing views in "their more simple expressions, in the belief that truth in philosophy, especially in areas of philosophy like the present one, is almost never very complicated" (1994 11). This has lead Strawson's opponents to attempt to formulate more complex accounts of how indeterminism might help with the problem, though it is not clear that any of these accounts can defeat Strawson's simple point. A further difficulty might be seen to lie in Strawson's definition of responsibility, which he phrases in a rather extreme way: "true moral responsibility is responsibility of such a kind that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven" (1994 6). Strawson points out, of course, that his argument is not meant to rest on this religious conception; rather, he is attempting to bring out the common notion of responsibility.

One problem, raised by Clarke (2005), points to the fact that, if we tone down the notion of responsibility to account for the fact that we are *finite* beings, and thus *eternal* punishment may be the wrong yardstick by which to measure our responsibility, we might end up with different conclusions. Another line of thought, raised by Ekstrom (2000), among others, points out that the question of moral responsibility is a

metaphysical question that is conceptually distinct from questions about the desirability, appropriateness, or justifiability of punishment. In some cases it may be entirely appropriate to hold someone responsible without punishing them at all; and questions about responsibility do, after all, make sense even if we believe that punishment, as such, is simply not the right response to violations of norms (a view defended, for example, by Sayre-McCord (2001). Smith (2005) goes one step further, pointing out that "being responsible" can and often does come apart from our reasons for "holding [someone] responsible." If so, not simply the issue of punishment, but the issue of whether or not we should hold someone responsible for their actions will simply be entirely irrelevant to the issue of whether or not one is, in fact, responsible. But it is not clear that these arguments serve to undermine Strawson's basic point. Making them stick would require a notion of responsibility for actions or character that is not merely independent of accounts of punishment and holding responsible, but that separates responsibility from the agent's control. And this is a far more difficult issue (I will address some attempts along these lines in the following section).

Clarke (2005) attempts, however, to tackle the metaphysical point of Strawson's argument directly, providing two lines of attack. Let me begin with the second. A slew of literature starting with Frankfurt's rejection of the Principle of Alternative Possibilities (PAP) (Frankfurt 1969) has attempted to establish that agents can be responsible for their actions regardless of whether or not they could have done otherwise. Frankfurt himself defended this claim by arguing that what matters for moral responsibility is that the first-order desire on which the agent acts be one with which the agent identifies through a second-order volition. If this condition holds, the agent is responsible for his action A

even if—counterfactually—some intervener was on hand to ensure that, had the agent in fact decided to perform some other action not-A, he would still have been forced to perform A. The moral of the story, then, is that the presence of the counterfactual intervener, and thus the agent's inability to do not-A, is irrelevant to the issue of moral responsibility given that the agent identified with his desire to A and did in fact A.

The literature on PAP is vast, and I do not have space to survey it here. But the lesson, contra Strawson, is meant to be this: even if an agent performs A necessarily because of the way he is mentally, this does not by itself detract from his responsibility. Thus, whether or not the agent is responsible for the way he is mentally is irrelevant to the question of his responsibility for A-ing. Clarke's point here is merely that Strawson's argument presupposes a premise—that if an agent performs act A because of the way he is, he can only be responsible for A if he is also responsible for the way he is—that a number of philosophers reject. In order to convince them, Strawson would have to show that their account of responsibility is mistaken.

No doubt my reply here will be taken as similarly insufficient. I have already, in Chapter 2, questioned the basic intuition behind Frankfurt's view of moral responsibility. To rehash: Frankfurt argues that agents are responsible, and act of their own free will, provided that their second-order volitions match up with their effective first-order desires. But Frankfurt considers irrelevant whether or not the agent is responsible either for his second-order volitions or for the correspondence between those volitions and their first-order counterparts. Whether or not an agent happens to be responsible, then, is a matter of luck. But this result seems counterintuitive: if I am not responsible for whether or not my action in fact stems from a desire with which I identify, responsibility is reduced to a

lucky coincidence: some people are responsible for their actions while others are not, just as some people are tall while others are short. But our responsibility does not, on this view, depend on what we do; it depends only on whether or not what we do also happens to be what we want to want to do. Should responsibility involve anything deeper, it seems, we must attempt to establish some responsibility for our attitudes, as well.

The other argument raised by Clarke emphasizes Strawson's reliance on the aforementioned premise that "what one does is a function of how one is." There are two ways of taking this claim: either what one does is *determined* by what one is, or it is not. Without offering any attempt at a knock-down argument, Clarke points out that Strawson's assumption that the latter is correct begs the question against a number of libertarians. Citing Van Inwagen and Nozick, Clarke brings up an alternative largely overlooked by Strawson's account. The alternative runs, roughly, thus: An agent might have reasons to do either A or B. That the agent has both of these reasons (or sets of reasons) is a result of how he is mentally. Assuming that the agent is not responsible for how he is mentally, it may still be possible for him to be responsible for whether he does A or B. If he does A, then he does it because of his reasons to A. If he does B, he does it because of his reasons to B. Thus, even though the agent is responsible neither for the reasons he has, nor for the fact that A is caused by one set of reasons and B by another, he may still be responsible for which set of reasons was in fact effective. All that is needed is that (1) how one is mentally does not *deterministically* cause the action and (2) the agent somehow determines which set of reasons in fact led to the action. Should this account succeed, agents could be responsible for their actions without having to also be responsible for their characters or attitudes or CPM.

The simplest response to this line is to press the standard objection to indeterministic accounts, as Strawson does in passing. Let us say that, given his CPM, an agent can do either A or B at time t and that whichever he does, A or B, it will be indeterministically caused by his CPM. This means that in some possible worlds, the agent will do A; in others, he will do B. In each possible world the agent has exactly the same CPM, but performs different actions. But surely, the objection goes, the agent is not responsible for *which* possible world he is in. But whether the agent performs A or B seems to depend entirely on precisely that; nothing about the agent determines what he does, since, by hypothesis, the agent is exactly the same in each of the relevant worlds right up to the instant when he decides on either A or B. And if agents cannot be responsible for the world they happen to inhabit, it seems they also cannot be responsible for which action they perform in the world they happen to be in. Clarke's point, of course, is not that the indeterministic model has a particularly devastating response to such arguments; his claim is only that, given the vast array of philosophical literature dealing with such problems, Strawson cannot simply assume the deterministic model. Should it be wrong, his argument will not be successful.

It is worth pointing out that the Basic Argument, like van Inwagen's Consequence Argument, is not a new argument at all. It is, as Strawson readily admits, as ancient as the problem of free will and moral responsibility itself. As with the Consequence Argument, in fact, it is difficult to see how the free will problem could be a problem at all for anyone who fails to take the argument seriously. These arguments merely express features of the phenomena of free will and moral responsibility, features without which these phenomena could not be what they are. One might thus wonder whether either argument

can be diffused through endless metaphysical quibbling over the nature of causation. There is, in any case, a further reason to take the view that responsibility for one's action requires a responsibility for how one is seriously.

Presumably, our actions have at least *something* to do with our characters. Perhaps, as the indeterminist claims, we might choose different actions given the same CPM, but deliberate actions do not simply flow from one set of attitudes or another. They are chosen on the basis of deliberation. And deliberation requires not simply having certain beliefs and desires, but also performing some operation involving those beliefs and desires, issuing in decisions, intentions, and actions. Nor are such deliberations typically simple: we rarely find ourselves in a situation with exactly two possible courses of action and exactly two, clearly defined, sets of attitudes supporting each one. What we encounter in deliberation is far more complex; our attitudes often remain shapeless and ill-defined before we reflect on them, and frequently do not gain much clarity despite entering into deliberative operation. The idea that desires and other pro-attitudes are clearly defined propositional attitudes is a useful simplification, but it can easily become misleading: in real life deliberation we frequently discover what we want, if at all, only when we first reflect on what we should do. Nor, incidentally, do we have well-defined algorithms telling us just which of our attitudes we must call up in making our decisions.

Part of the messiness of deliberation is recognized by Davidson and offered in his explanation of weakness of will. (Davidson 1980d) The scope of beliefs and desires involved in causing our actions may well be far narrower than the scope of beliefs and desires that enters into forming our all things considered judgment concerning the best course of action. Moreover, as Arpaly (2003) adds, there is no guarantee that the all

things considered judgment agents come to in deliberation will, in fact, consider all relevant things: we are liable to miss all sorts of relevant desires and beliefs, precisely because we lack the perfect algorithm for gathering an exhaustive list. (Unfortunately, Arpaly equates the actual judgment an agent reaches with the agent's "best judgment," though it is fairly clear that, on any common meaning of that term, the point is instead that the agent's considered judgment may well fail to be her best judgment.) Advising agents to follow their considered judgment will not, then, guarantee that their action best represents their character; nor, obviously, will advising them *not* to follow that judgment.

Arpaly herself uses the argument as a springboard for the further insistence that neither our rationality, nor the blameworthiness or praiseworthiness of our actions, depends on the extent or even the presence of deliberation in producing our actions. I will return to that point. For now, however, I want to raise the following issue: What we decide to do depends on (1) how clear we are about the relevant character states, (2) how clear we are about which character states are relevant, and (3) which character states finally cause our actions. But all of these features are, themselves, products of character. Some people are better attuned to their wants and beliefs than others; some are better at finding and considering the relevant ones; and some are better at having their actions stem from the best considerations.[26] But how good agents are at each of these tasks is a matter of character, and insofar as one's CPM is behind the clarity with which their choices are made, and the efficiency with which the clearer choices determine one's

[26] I do not mean to refer here, exclusively, to continence in the classic sense as involving strength of will. Having the sort of character that allows one to follow through on one's judgments of what is best is one thing; having the sort of character that facilitates having the rights sorts of mental connections—connections that link decisions about what to act to decisions about what is best—is another. We often do what we should not merely through inattention; whether one wishes to call such cases instances of "weak will" is not important to my considerations here.

action, character clearly plays a crucial role in the production of action, and one that does not—in any obvious way—allow the final choice of action itself to ground the agent's responsibility.

We can reach a similar realization if we follow Strawson's own insistence that thought is not a kind of action. "The role of genuine action in thought is at best indirect. It is entirely *prefatory*, it is essentially—merely—*catalytic*." (Strawson 2003 231) While we may actively decide to think about a certain topic, keep ourselves from wandering away from it, or try to reinforce the thought process in various ways, there is "no action at all in reasoning and judging considered independently of the preparatory, catalytic phenomena just mentioned" (2003 232). So whatever role action plays in getting us to think and keeping us there, "the coming to mind itself—the actual occurrence of thoughts, conscious or non-conscious—is not a matter of action." (2003 234). And this view, that thought is something that "just happens" to us, can be carried into the sorts of thoughts that constitute our judgments about what to do and even our decisions, so that "most deciding what to do is best seen as something that just happens, even if there is also, and crucially, some sort of genuine action of positive commitment to the decision, either at the time it is reached, or at the moment of the 'passage à l'acte'." (2003 244)

So while our actions are up to us, in the sense that we can decide what to do, the decisions themselves do not seem to be up to us in that way any more than the content of our deliberations—what we actually happen to think about when reaching our decisions—is up to us. It is, incidentally, for this very reason that virtue ethicists tend to stress the moral importance of being a certain kind of person: if our chosen actions flow from the thoughts that occur to us, then being the sort of person to whom particular

thoughts and not others occur enters into considerations of moral responsibility. Whether or not immoral considerations occur to us in the course of deliberation may have only a statistical correlation—if that—to whether or not we perform immoral actions. But whether or not *moral* ones do will certainly make more of a difference. Moreover, whether we decide to act on our moral or immoral considerations, provided that both occur to us, may well be a matter of which way our deliberation happens to go and the idea that we *actively* choose a course of action, or make a judgment about which course of action is best, is just what Strawson is questioning.

One can, of course, take note of Strawson's use of "most" in "most deciding what to do." And this might open the way for a restrictivist response: while *most* of our decisions are the sorts of events that "just occur" to us, some are not, and perhaps we can hold *those* to be the locus of responsibility. But there is a sense in which Strawson's "most" is just a way of being agreeable: ultimately, even if some of our decisions are actions rather than just thoughts that occur to us, those decisions themselves will need some preconditions in thought. In other words, even if some of our decisions involve agency, they do not *ultimately* involve it: we are not *causa sui* with regard to anything we decide.

Once again, then, we come back to character. The attitudes or character traits that cause our actions do, of course, belong to character. But so do the other attitudes, the habits of thought, of seeing relevance, of attentiveness, and so on. Discussions of character and of attitudes frequently focus only on the narrow, former group. But it is the latter group that forms the decisive links between what we are mentally (in the narrow sense) and what we do. Our character does not merely provide us with the raw materials

for taking actions; it serves also as the inescapable *background* of those actions. With this acknowledgement I want merely to complicate any position that attempts to make our responsibility for our actions independent of our responsibility for character, for how we are mentally. Indeterminism or not, the decisions that lead to actions rise out of a complex soup, much of which we are not aware of; as the term "background" suggests, much of it we cannot be aware of insofar as we are engaged in making decisions. If, then, we are to be responsible for our actions, it is difficult to avoid the conclusion that we must also be responsible for the background of those actions, for how we are mentally. With this consideration in mind, I now turn to recent discussions about the nature of such responsibility.

B. Responsibility and Consciously Thematized Decision

While Frankfurt has changed his mind about what is required for identification with a desire, a constant theme in his work has been that responsibility requires a match between certain of the agent's attitudes. Depending on how one explains identification, then, it becomes possible to give an account of responsibility that is detached from the agent's voluntary control over her actions or attitudes; what matters for responsibility is whether an agent's action is representative of her Real Self.[27] This theme has in recent years been combined with a view of responsibility developed from Peter Strawson's influential "Freedom and Resentment" (1962), which appeared nine years prior to Frankfurt's seminal work. In that article, Strawson proposes that we accord priority to the reactive attitudes—attitudes such as resentment, gratitude, anger, and so on—and

[27] This label originates in Susan Wolf's work.

disconnect the practical questions of whether we can or should give up such attitudes from the theoretical question of whether or not agents do in fact have metaphysical free will. If we should and must treat someone *as if* they are responsible, the suggestion runs, then they *are* responsible for all practical purposes.

Like Frankfurt's work, then, Strawson's account attempts to free discussions of responsibility from metaphysics. But it adds two other crucial (and connected) components as well. First, it centers on our actual practices of praise and blame, focusing specifically on what about the agent evokes our reactive attitudes. Second, in answering this question, it proposes that we feel gratitude or resentment, anger or sympathy, on the basis of the quality of the agent's will. Thus, for example, we will resent someone who acts out of malice while we may not resent—or may resent less—someone who carries out the same action out of a mistaken sense of love or loyalty to a worthy person or cause. And we may not resent at all someone who acts out of a compulsion, especially one that runs counter to her overall attitude. The combination of these two approaches—the Real Self View and the quality of will theory—has led to a number of new theories of responsibility. In Chapter 5, I will address one of these—Fischer and Ravizza's semi-compatibilism. Here I will focus on another: attributionism. By contrasting it with the volitionist view of responsibility, I will attempt to shed some light on the notion of control required for responsibility; in the next chapter I will look at the notion of choice.[28]

[28] I take the terms "attributionism" and "volitionism" from Neil Levy (2005). Levy borrows the term "volitionism" from Smith (2005), who refers to her own approach as the "rational relations" view and groups it under the general category of "non-volitional" accounts (2008). I will use Levy's terminology here.

Attributionism is the view that an agent is responsible for her action, omission, or attitude[29] if it can be attributed to her as an agent.[30] Volitionism, on the other hand, takes the more traditional line that we can be responsible only for something we have chosen and have control over. Or so the standard definitions go. The central—and to attributionists objectionable—aspect of volitionism is the claim that an agent can only be responsible for that which is traceable back to his choice, voluntary control, or conscious deliberation. But this rough characterization obscures what is, on my view, the central bone of contention between the two accounts: by presupposing that we already know the meaning of terms such as "deliberation," "choice," "control," and "voluntary," the parties to this debate seem to me to conceal the central requirement for responsibility. What requires clarification—and what is really at stake in the debate—is not whether responsibility requires something like choice or control, but rather what *kind* of choice or control is required. And the problem for theorists of responsibility, I believe, is to develop an account of these notions that allows attributionism and volitionism to converge.

I will develop this thought as I go along. For now, however, I want to make one further terminological claim. "Conscious" and "unconscious" are frequently used in a haphazard way, so that the descriptive "unconscious" is made to cover everything of which the agent is not aware, from deep-seated drives to processes occurring just below the level of awareness. This terminological haziness forces consciousness into a thin skin,

[29] Smith specifically defines her view as one that accounts for responsibility for attitudes and—she thinks—omissions. Arpaly and Sher take in actions as well.

[30] On Smith's view, "according to these philosophers, what really matters in determining a person's responsibility for some thing is whether that thing can be seen as indicative or expressive of her judgments, values, or normative commitments." (Smith 2008 368) While this is a fine summary of her view, however, it is too narrow to encompass all the views she has in mind. Sher, in particular, argues that this account is still too narrow to account for the entire range of our ordinary judgments of responsibility.

with only enough room for thoughts to which we have immediate, effortless access. It thus problematically drives the vast majority of our thoughts, decisions, and actions into the realm of the unconscious, a realm for which—as the dominant paradigm would have it—we lack responsibility. To avoid prejudging the issues in this way, I will treat the disagreement between attributionists and volitionists as centered not on choice, control, voluntariness, or conscious deliberation, but on consciously thematized decisions (CTDs). These are decisions or choices that agents make not merely with awareness of what they are doing, but with a thematized awareness—such a decision (and perhaps the deliberation leading up to it) is, so to speak, the focal point of the agent's awareness at the time it is made. The debate may thus be rephrased in the following terms: volitionists believe that responsibility requires CTDs; attributionists do not. But what, then, do attributionists think responsibility requires?

Here I will look at three approaches, developed by Angela Smith, Nomy Arpaly (alone and together with Timothy Schroeder), and George Sher. These theorists share the Frankfurtian view that responsibility for a thing requires some straightforward connection between that thing and the self. At the same time, they reject Frankfurt's insistence that the "self" relevant to responsibility consists of higher order volitions, by means of which agents either identify or fail to identify with their first order desires. Consider, for example, the case of a man who knowingly treats others badly, although he does this unwillingly. Surely, we might say, his unwillingness to be a bastard is not, by itself, enough to excuse him from responsibility. This is even clearer if we take a brief look at the version of the Real Self view developed by Gary Watson. Concerned that Frankfurt's account of identification allows for an infinite regress of higher order volitional states,

Watson suggested that we contrast desires and values; our Real Self, on this view, is to be identified with our values (Watson 1982). But this hardly works any better: when a person acts contrary to his values, this does not—intuitively—excuse him from responsibility; rather, it compounds it: not only does he act wrongly, but he betrays his values while doing so. Responding to such concerns, attributionists seek to link responsibility to a notion of self far broader than identification, choice (in the volitionist sense), or reflective endorsement. The goal—and this is the aspect of attributionism that will particularly interest me—is to show that these intermediaries are not needed in an account of responsibility. Instead, the idea goes, we can be responsible for actions and attitudes because these can be immediately and directly attributed to our selves.

A further motive for the development of attributionist theories lies in taking seriously the Strawsonian focus on our actual practices of praise and blame. In ordinary, everyday judgments of responsibility we do not, as a matter of practice, seem to focus on the agent's choice. Or, at least, it is clear that if we *were* to do so, many of our ordinary judgments would turn out to be false, and we often attribute responsibility for actions and attitudes that were clearly not chosen by the agent, a point illustrated in detail through the many colorful examples drawn on by attributionists. Traditionally, volitionists have attempted to deal with the problem through tracing, i.e., the view that in cases where agents seem to be responsible for an action or attitude, and yet clearly have not chosen it, we might show them to be responsible by tracing the development of the attitude or the creation of a situation in which the action became unavoidable to some earlier act of choice on the agent's part (theories of tracing often draw on updated versions of the Aristotelian account of habituation). Though such an approach will—in most of the

problem cases—seem implausible to anyone not in the grip of a particular theory, it has recently come under sustained theoretical fire as well. Vargas (2005), for example, presents a series of clear examples such as that of Jeff the Jerk. Jeff has the job of firing people, and fulfills this job in the rudest possible way because he is, to put it simply, a jerk. He is, intuitively, responsible for the insensitive way he treats those he fires. But, as Vargas neatly shows, it is fairly easy to construct a back story on which Jeff does not, at any point in his life, knowingly make choices that will lead him to become the sort of person who will treat people badly as he fires them. In fact, it would be rather difficult, in cases of this sort, to construct a different back story. No doubt we can sometimes choose to cultivate particular sorts of habits and character traits, and we might even choose to do so knowingly, but this sort of deliberate character formation accounts for only a tiny minority of the character traits of normal human beings.[31] If we are to account for what seems like a vast majority of our ordinary judgments, then, a volitionist view once again seems insufficient for the task.

Smith develops a view clearly meant to handle these two difficulties. To be proper objects for moral appraisal, a person's actions, attitudes, or omissions need have no particular history, nor must they stand in a relationship removed from the agent by some volitional act of choice or identification. Rather, they need to directly "reflect her

[31] Vargas's point is even stronger. He argues not simply that Jeff did not know that some of the actions he undertook as a teenager would lead him to become a jerk, but that Jeff certainly did not know that whatever character-forming practices he engaged in as a teenager would lead him to one day be insensitive in firing people. I would venture that there are two further epistemic criteria for "choice" in the volitionist sense that character-forming acts will fail to meet, both resting on the impossibility of knowing the future. First, since we lack anything like precise knowledge of how character-formation works, we cannot have full knowledge of how even the most deliberate character-forming acts will shape our character. Second, since when I choose acts that I believe will form a certain character, I cannot, in principle, know what it will be like for me to have that character. I can know what it is like to act *as if* I were a jerk, but I cannot know what it is like to think and act *as* a jerk since, presumably, I have not yet managed to form that character trait. Even if I choose to become a jerk, then, I cannot do so in the full knowledge of the sort of character I am, in fact, acquiring for myself. (For example, I may assume that I will still have certain options of thinking and behaving open to me that will, in fact, never even occur to me once I become a real jerk.)

practical agency." (Smith 2008 381) What counts as belonging to a person's practical agency, in turn, must be broad enough to include the sorts of things we commonly hold people responsible for, yet narrow enough to exclude inexplicable urges, implanted thoughts (such as those considered by various "manipulation cases"), and physical features (which, though they may play some role in agency, do not belong to agency as such). What allows us to delineate the field of responsibility in this way is an appeal to rationality: agents are responsible for those things that, roughly speaking, have a rational content and are at least somewhat integrated into the agent's overall normative framework. Since physical features have no rational content, whereas inexplicable urges and implanted thoughts will lack any normative connection to the agent's other beliefs and judgments, this criterion seems to contain just what is needed while excluding anything external to agency; this view therefore "gives us a satisfying account of the boundaries of the moral self." (Smith 2005 263)

In working out her account in a plausible way, Smith discusses the sorts of things we hold people responsible for, showing the ways in which they reflect evaluative judgments on the part of the agent. In particular, Smith emphasizes that we often hold people responsible for things frequently left out by standard volitionist accounts: patterns of noticing or neglecting states of affairs around us, the sorts of thoughts that occur to us, and even some involuntary reactions. Consider the first case: it is one thing for a driver to notice a child riding a bicycle near his car, and whether or not a driver sees this may not reflect his practical agency in the least. But it is quite another thing for a driver to take care to keep track of the child's position in order to minimize any risk of a collision. A driver who fails to continue to keep tabs on the child, or perhaps does not notice that the

child is in poor control of his bicycle, is blameworthy for this oversight. Someone concerned for the child's safety, on Smith's reasoning, *would* notice the child's lack of control, although these activities of noticing and monitoring need not be consciously willed. It is simply the case that drivers who hold the safety of children to be important notice certain things that other drivers do not, and the difference between these drivers is a morally relevant one. The driver who fails to take further notice of the child is blameworthy not because he has decided not to take further notice, or even because he at some prior point decided not to care about children, but because he does not care now, or does not care as much as, perhaps, he cares about getting to his destination on time.

Something similar is at work in the patterns of what occurs to us, quite involuntarily. Smith illustrates the idea with the example of a businessman who wonders whether having a rival killed might solve his problems. The businessman has no voluntary control over whether such a thought occurs to him and, furthermore, he may immediately dismiss it. Yet the fact remains that there is a difference between people to whom such thoughts occur and those to whom they do not. Of course the businessman who actually puts out a hit on his rival is far more blameworthy than one who merely considers it, and one who has the thought but dismisses it out of hand is less blameworthy still—hardly at all. But Smith, like the other attributionists, works hard to distinguish responsibility from blame or praise: to be open to attributions of responsibility is not, necessarily, to be deserving of blame or praise; in fact, there is something ethically suspect about the person who insists on blaming someone for having a fleeting thought. (Smith 2007) Yet having such a fleeting thought reveals a flaw in one's character nevertheless.

Finally, involuntary emotional responses are generally thought to be entirely free from moral evaluation. Smith contends, however, that like the previous cases, they can show something morally relevant about the agents who have them. Someone who gets annoyed when asked to donate to a cause, for example, is someone who likely is not committed to that cause; and someone who does not regret forgetting a dinner date with a friend likely has no particularly warm feelings toward that friend. Our involuntary responses—despite obviously not being consciously chosen—can thus reflect our underlying commitments. This does not, of course, mean that all voluntary responses, or all failures to notice something, or all fleeting thoughts, really *do* reveal something about our deeper underlying evaluations. As Smith admits, we can have islands of irrationality, as an agoraphobic's fear of heights might persist despite her judgment that the railing in front of her is perfectly safe. Nevertheless, much of the spontaneity of our thoughts, emotions, and reactions is reflective of rational commitments that support these responses. Importantly, if this account of responsibility is correct, it can explain why we hold people responsible for despicable attitudes, various omissions, and even actions that clearly occurred in the absence of deliberation. Moreover, in light of the earlier discussion of the Basic Argument, attributionism can provide an explanation of why we hold people responsible for clearly voluntary and deliberate actions despite the fact that the considerations that enter into those deliberations are not, in the volitionist sense, within our control.

What makes us responsible for attitudes on this account, then, is that they reflect underlying judgments on the agent's part, either by directly embodying those judgments, as a contemptuous attitude toward members of a particular race embodies a judgment of

their inferiority, or by standing in a rational relation to them, as a fear of the Roma might depend on an underlying judgment that they are likely to pick one's pocket. These underlying judgments need not, as Smith stresses, be ones that the agent has reached through conscious and explicit deliberation, or CTD; they may well be "things we discover about ourselves through our response to questions or to situations" (2005, 252). They may, in other words, belong to the background operating behind the scenes of an agent's deliberations, perceptions, emotional responses, and so on. And what makes it appropriate to hold someone responsible for such judgments or the various states that reflect them is that they are, taken together, constitutive of a person's practical agency. "We are not merely producers of our attitudes, or even guardians over them; we are, first and foremost, inhibiters of them. They are a direct reflection of what we judge to be of value, importance, or significance." (Smith 2005 251)

In moral appraisal of a person, then, we do not simply provide a negative description of them, as we do when we comment in unpleasant ways on someone's weight, hair color, or inability to perform simple mathematical tasks (unless, of course, that inability reflects a judgment that the tasks at hand are not worth bothering with). Instead, we address a *demand* to the agent. We ask the agent to justify her attitude or the underlying judgment, to "explain or justify her rational activity in some area, and to acknowledge fault if such a justification cannot be provided." (Smith 2008 381) This need not mean that they must reply, of course; as Smith notes, "this is not to say… that I will necessarily acknowledge or take seriously such a challenge; it is a point, rather, about the nature of moral appraisal itself, and how it differs from mere negative description." (2008 381) A further point, one which seems to follow but which Smith

unfortunately does not address, is that agents need not be *able* to provide anything resembling a justification. Rather, the point about the legitimacy of a demand for justification is intended only to delimit a class of things for which someone may be held responsible, in much the same way as the legitimacy of the question "why?" may delimit the class of intentional actions, in Elizabeth Anscombe's famous formulation. (Anscombe 2000 11)[32]

This last point, about the agent's ability, strikes me as important. Moral philosophers too often seem to think of adult human beings as universally endowed with the ability to defend and justify their views, and it is a source of wonder that, despite the constant challenges to this view presented by daily interaction with students (not to mention the hordes of non-philosophers we tend to bump into in the real world), recognition of this fact so rarely seeps into writing on rationality, agency, or ethics. But ordinary people—as well as most philosophers—are not particularly good at justifying their moral attitudes; if they feel pressed to offer a justification, the justification is frequently ad hoc, and likely not reflective of their actual judgments. This recognition, I think, should make Smith's point a bit more radical than she wishes it to: the legitimacy of a demand for justification is the mark of that for which one is responsible; but it is so regardless of whether the agent can provide any such justification. We are, in other words, responsible for our attitudes insofar as we can be asked to justify them, not insofar as we can actually provide the justification.

[32] This raises an interesting suggestion: that the difference between an intentional action and an action for which an agent is responsible will revolve around the difference between the question "why?" and the demand for justification. Whether we are responsible for all our intentional actions, then, will depend on whether we take all motivating reasons to also be justifying reasons.

The point here is supposed to be one about the connection between responsibility and agency. The demand for justification "*by its very nature* implies responsibility, for it is directed at [the agent's] judgmental activity, activity for which we must regard him as responsible if we are to regard him as a moral agent in any sense." (Smith 2008 388) What makes responsibility judgments possible, then, is—as in the volitionist case—the *activity* of the agent, but the activity need not be conscious, explicit, voluntary. This account, then, is meant to leave control (in the volitionist sense) entirely out of judgments of responsibility, and this has led some to attempt to soften the blow rather than reject the theory outright. Michael McKenna, for example, suggests that control has to be involved somehow, but that it has two components, one of which is Smith's, and "involves the possibility of rational activity (that, let us grant, falls shy of free mental acts). A second involves a standing capacity to perform a free mental act of deciding or choosing to evaluate one's moral standpoint(s)." (2008 36) Even this much control, however, is too much.

Consider Smith's insistence that we can hold someone responsible "even if the person's failure was not a failure of choice, and even if she is not in a position to change her attitude 'at will.'" (Smith 2008 383) Of course this is not an explicit rejection of McKenna's second condition; after all, an agent may have the capacity to *evaluate* without being in a position to change an attitude at will. But *such* a capacity fails to add any control whatsoever. A weaker version of the second condition is, in fact, built into Smith's account, since for someone to be responsible for an attitude, "it must be the kind of state that is open, in principle, to revision or modification through that creature's own process of rational reflection." (Smith 2005 256) That a state is "open" to revision in this

way does not imply that the agent can simply decide to change it and thereby make it so; it indicates only that the state belongs to the agent's rational activity, and whether or not it changes is, thus, attributable to the agent qua agent. In any case, an attributionist will have reason to doubt whether McKenna's second condition is even coherent in light of his other concessions; whether or not an agent performs "a free mental act" must, after all, itself depend on whether the agent *judges* such a performance to be worthwhile. Should we then have to look for an extra capacity to evaluate (and change) that judgment itself—as the idea that "control must come in somewhere" seems to demand—we are stuck in an infinite regress. It seems as if either one accepts attributionism or rejects it, but combining it with volitionist choice at any level leads nowhere.[33]

As I mentioned earlier, a driving motivation for attributionism is the perceived failure of volitionist and Real Self theories to adequately delineate the boundaries of moral agency: an agent who acts against her best judgment or on the basis of a desire with which she does not identify is not, intuitively, exempt from responsibility on that ground alone. Overcoming the limitation thus drives an expansion of the boundaries toward a more holistic view of the agent. Responding to standard philosophical challenges of manipulation by brainwashing masterminds or evil scientists, for example, Smith articulates the difference between implanted attitudes and ones that are the agent's own. Only the latter are based on judgments that reflect the agent's practical self, and we see this by comparing them with other judgments that the agent holds to find patterns; implanted judgments and attitudes do not fit into those patterns. We thus figure out what attitudes are the agent's own by seeing how they fit into the overall framework of

[33] I do not mean to suggest that no rapprochement between volitionism and attributionism is possible; on the contrary, I will argue that it is necessary. My point is only that whatever control condition one wishes to read into an attributionist account cannot be a expressed in volitionist terms.

attitudes and judgments on the basis of which we attribute responsibility. And this appeal to holism is, Smith argues, a strength of her account, since a good account of responsibility "should preserve our sense of the rational interrelations among our attitudes, rather than treating these things as isolated entities, each of which must meet some further criterion before it can be considered attributable to a person for purposes of moral assessment." (Smith 2005 262) A person's rational agency is not a collection of unrelated items; they are items rationally bound together into a web, that web being the agent's moral self, and the binding her rational activity.[34]

The goal, thus, is to develop a conception of the self that does not restrict its "real" or essential characteristics to some volitional core. Arpaly and Schroeder take up this challenge with "the Whole Self View." Unlike the Real Self View, which looks at whether the desire behind an action is linked to some higher order volitional state or value, the Whole Self View aims to take into account the myriad volitions, attitudes, and beliefs of an agent. A kleptomaniac who acts unwillingly and against all his desires may, on this view, be excused from responsibility. But what of the kleptomaniac who, looking back on his petty thefts, inwardly smiles to himself? Human beings have complex relationships to their actions, and this complexity is not exhausted by pointing at some simple feature such as identification with a volition.[35] Accordingly, the Whole Self theory

[34] I leave aside the issue of whether it still makes sense to speak of "states" in this regard. The idea of a "state" implies a stability that may not sit well with the ongoing, constantly unfolding process of an agent's rationality at work, binding together her various judgments and attitudes into a more or less coherent whole. But I mention this point to suggest that the reference to a "state" here is theoretical; in theory and in explanation we reconstruct the agent as having "states" due to the difficulty of referring to transitional processes as the causes of action.

[35] Frankfurt hints at some recognition of this in later work, where he notes the phenomenon of ambiguity—where we are torn between identifying with a desire and identifying with its opposite—and suggests that such ambiguity may not be completely avoidable for human beings. If so, the Real Self View ultimately collapses into the Whole Self View.

links responsibility for an action to the degree to which "the morally relevant psychological factors underlying it are integrated within [the agent's] overall personality." (Arpaly and Schroeder 1999 172) Those factors—beliefs and desires—are taken to be "well-integrated within a person to the extent that (1) they are *deep*; and (2) they do not *oppose* other deep beliefs or desires"; an action, in turn, is well-integrated "to the extent that it results from such beliefs and desires." (1999 173) Depth, in this case, is understood as the force which the belief or desire in question has in the person's agency—how likely it is to influence actions, and how difficult it is to overturn it by other considerations. And the less opposition there is between the mental states in question and other—rationally incompatible—ones determines their importance to the agent as a whole. The condition of responsibility, then, is not some mental state or act that legitimates a trait as the agent's own; this purpose is served by the entire nexus of the agent's character.

This view lies in the background of Arpaly's *Unprincipled Virtue*, which aims to overturn accepted theories of rationality and moral worth. Drawing on a host of literary and real life examples, Arpaly argues that an agent's deliberately reached conclusions and consciously held principles may well be at odds with the agent's character as a whole, and this fact should lead us to reexamine the reliance of moral psychology on accounts that inevitably privilege the side of the conscious and explicit. With regard to moral worth, she draws on the example of Huckleberry Finn (among others) to demonstrate that agents may act in praiseworthy (or blameworthy) ways despite failing to act on their consciously held principles. Huck believes that it is wrong to help slaves escape, and he helps Jim in spite of that belief. If—as Arpaly thinks our intuitions

demand—we are to take Huck's action as praiseworthy, we should recognize that what makes it so is not any reason Huck happens to hold explicitly, but just the opposite. Huck helps Jim because he responds to Jim's humanity, recognizes Jim as a fellow person—though not, to be sure, in those terms—and what makes Huck's action praiseworthy is that these reasons are precisely the reasons that (objectively) make it right to help Jim. Arpaly's account, then, holds that agents are morally praiseworthy insofar as the reasons on which they act, perhaps unbeknownst to them, are the reasons that make their action right; they are blameworthy when the reasons on which they act are ones that make the action wrong; the depth of the agent's commitment to the right (or wrong)-making features determines the degree of praiseworthiness or blameworthiness.

That some of the attitudes in an agent's repertoire happen to be consciously held, deliberately chosen, or voluntary is, for the most part, irrelevant, and the same account plays out in Arpaly's view of rationality. An entire chapter is devoted to defending the possibility of reverse akrasia, in which an agent acts rationally despite acting contrary to her own best judgment.[36] Reverse akrasia, Arpaly insists, is precisely what we find in Huck's case, since he acts contrary to his own explicit judgment and yet does what she—and, she assumes, her readers—holds to be the rational conclusion. What it is *rational* for an agent to do is to act on the reasons best supported by the agent's mental framework as a whole. Deliberation and reflection may help us to discover those reasons, since they serve to "*focus your concentration*, allowing you to pull together mental resources from many different corners of your psyche in order to solve whichever problem you have decided to reflect on." (Arpaly 2003 64) But they may also fail to help, since deliberation

[36] The phrase "best judgment" is somewhat misleading. What Arpaly means by this is the conclusion that an agent reaches after explicit deliberation; on a more common-sense use of the term, Arpaly's view is precisely that deliberation may result in the agent's reaching a quite poor judgment.

itself may well be distorted by self-deception, wishful thinking, or some other irrational condition. In Arpaly's example, Sam, a student, believes that in order to get his studying done, he must stop interacting with people and become a hermit. Although he reaches this belief through deliberation, however, the belief is mistaken: Sam ignores a good deal of evidence that suggests, for example, that he works much less efficiently when distanced from other people. Thus, while becoming a hermit may be the course of action recommended by Sam's deliberation, and acting contrary to this judgment is an akratic act on Sam's part, the truth remains that, given his goals, continuing to interact with others is the rational course of action for Sam to take. The rational course of action is determined by reference to the agent's character, the whole of his desires, beliefs, and attitudes; since agents lack access to the whole of their character in deliberation, they can easily reach mistaken conclusions about what the rational course of action for them is.

To further illustrate the claim that rationality does not require deliberation, Arpaly points to common cases of rationality that do not involve it, at least in the standard sense. A fast-acting athlete, for example, reacts instantaneously to a change in the play; there is no time for deliberation, and yet the reaction can be a rational or an irrational one. Of course someone might deny that fast-acting athletes act for reasons at all (Dreyfus, as we will see, makes precisely this claim), but Arpaly suggests that this runs counter to our intuitions. "A major part of what it is to be a competent tennis player is to have the ability to play tennis rationally—to act for good reasons rather than bad reasons in all of your game-related actions." (2003 53) If a jump to the right will allow a player to hit the ball while a jump to the left will not, the player has a *reason* to jump to the right. One could reply that even though the player has a reason to jump to the right, he is not acting on a

reason when he does so: he is merely acting on instinct. But this could be a hard sell: as Arpaly's remark about competence suggests, a player trains precisely to match his instincts to the right reasons, so that in acting on instinct he *is* acting for a reason. If he were not responding to reasons at all, it would be hard to make sense of the fact that the expert's instincts are so closely aligned to reasons.

In two further cases, Arpaly notes experiences of having a realization dawn one one as, for example, a man might realize that he is in love with his childhood friend without having come to this realization through deliberation. The thought suddenly occurs to him, and occurs with an obvious sense of its truth; and it may well be a realization that has been long in coming, and has involved a good deal of sub-conscious responsiveness to reasons—to this evidence of his actions and reactions where his friend is involved, say—over a long span of time. Finally, deliberation itself can be a rational process without involving further deliberation to support each step in thought. That would threaten an infinite regress and, in any case, Arpaly notes that emotional responses play an important role in deliberation: one can quickly move from one step in the deliberative process to another because a particular thought feels right (the earlier discussion of Galen Strawson has already addressed these issues). What the theory of rationality suggested here implies, then, is that agents can and frequently do act for reasons without knowing what those reasons are and certainly without explicitly recognizing that they are acting for reasons. A match between the reasons for which the agent acts and the actual reasons to do something is all that is needed for an action to be a rational one; the agent's own attitudes or beliefs with regard to her reasons are largely irrelevant. Whether or not my attitudes and beliefs on balance favor a certain course of

action determines whether the pursuit of that action is rational; whether or not some of those attitudes or beliefs are conscious ones is irrelevant.

Against this backdrop, Arpaly presents her account of moral responsibility, the features of which should already be clear: we are responsible for actions, i.e., open to the possibility of praise or blame for them, if those actions are taken for reasons supported by well-integrated attitudes. Once again, then, the argument is clearly an attributionist one: the agent's choice, deliberation, and volition do not come into play in attributions of responsibility; Huck is responsible (and thus open to praise) for helping Jim because the reasons for which he helps Jim are good ones, and because they are integrated with the rest of his character. Arpaly concedes that Huck would be even *more* praiseworthy if he did not have, as part of his character, the belief that helping slaves escape is wrong. But the fact that Huck acts contrary to what he *thinks* are good reasons, and instead acts for what he does not even recognize as reasons does not detract from his responsibility. Similarly, Arpaly draws on Le Carré's character of Oliver Single, who betrays his criminal father to the police without ever choosing, or deciding to do so in order to make the point that choice, or CTD, is neither the necessary nor a sufficient condition for attribution of responsibility.

Sher's account differs slightly from the others, but begins in a similar vein, by enumerating case after case in which we are likely to hold an agent responsible despite anything like deliberation, choice, or control (in the volitionist sense) being present. Among his examples: Alessandra, who leaves her dog in the car on a hot day while picking the kids up from school but forgets about the dog when she is approached by the principal about her children's performance; Joliet, home alone, hears movement

downstairs and, fearing it to be a burglar, gets her gun and sneaks down; seeing a figure, she panics and shoots, but it turns out that she has shot her son. These—and many other characters—are taken as typical examples of agents whom we would intuitively hold responsible; clearly, however, volitionist control is missing in these cases. The characters act on poor judgment or get distracted, and do so in culpable ways, but they certainly do not choose to exercise poor judgment or to become distracted—clearly, had they exercised any explicit judgment at all, most of them would have avoided the blameworthy action. The argument, then, is that if we think of control as involving a conscious awareness that one is acting wrongly, we frequently hold people responsible for actions beyond their control and a theory of responsibility should account for this.

Sher's reply takes an unusual form. Rather than insisting that control is unnecessary for responsibility, he suggests that we redefine control in a non-volitionist way. Rather than thinking "that an agent's control extends no further than the searchlight of his conscious awareness" (Sher 2006 296), Sher suggests that control may be a product of the agent as a whole person. Agents have "innumerable beliefs, desires, motives, convictions, and commitments of which [they are] not aware" and these may well be the agent's own just as much as his conscious attitudes. If the failure to recognize that the action is wrong is itself caused by some combination of such attitudes, this may yield the sense of control we are looking for. Thus, for example, Alessandra would not have forgotten her dog had she cared less for her children and more for the dog; and she has control in the sense that the attitudes she *actually* has, in the strength and combination in which she has them, caused her to neglect her dog. With Joliet, the case is harder, but "we can say that the tendency to panic that prevents Joliet from recognizing that she

should not pull the trigger is itself part of what makes her the person she is." (Sher 2006 301)

Despite lacking an account of what it means to say that the attitudes in question are the agent's *own*, an account it seems we may have to fill out with some view of holism or integration, Sher's argument is quite close to the attributionist view. In fact, his one disagreement with the attributionists is, I think, unwarranted. Sher notes that on the attributionist view, the wrongful act must be caused by some combination of the agent's states but that, furthermore, there must be an "ineliminable semantic component" present, since "this account makes essential reference to the match or fit between the relevant feature of the act and the contents of the attitudes or judgments that determine the agent's practical identity." (Sher 2008 225) His own view, Sher insists, has no such implication and consequently can apply to a wider range of cases. Specifically: "If we accept the attributionist account, then we will have to attribute Alessandra's responsibility to a judgment that [her dog's] safety doesn't matter much, or to a lack of good will toward the dog, and we will have to attribute Joliet's responsibility to a judgment that it is not important to take precautions against inflicting serious harm." (Sher 2008 226) But even if such judgments are absent, we may still hold Alessandra responsible for neglecting her dog's safety, and "Joliet would surely remain responsible if her panic had simply overwhelmed her judgment." (2008 226) Our judgments of responsibility here, then, do not depend on the semantic connection.

But this seems mistaken, at least when taken at face value. First, Smith, at whom the remark seems to be directed, clearly does not believe that there must be a direct match between the act and a judgment of the agent. Her discussions of responsibility for what

we notice or miss, what occurs to us, and so on, would make no sense if that were so (the businessman who considers having his rival killed is blameworthy—on her account—not because he judges that killing business rivals is good, but because he seems to lack commitment to the judgment that such actions are impermissible). Nor would Smith's central cases of omission—forgetting a friend's birthday, for example—make sense on Sher's description. Second, Sher's own account of what gives Alessandra control draws on the same considerations employed by Smith. Third, in a footnote[37] Sher implies—though he does not say so—that the connection between the agent's attitudes and her action that he has in mind must not merely be causal, but must also be appropriate (otherwise his account would seem too bizarre). But giving sense to the notion of appropriateness without any semantic component might prove difficult. Fourth, drawing on the last point, we might ask whether a *merely* causal account could make any meaningful sense of Joliet's responsibility. There is no way whatsoever in which panic—taken in isolation—can appropriately cause any action whatsoever. And, moreover, if we did genuinely believe that Joliet was overtaken by panic and that was all there was to the story, we could not hold her responsible. We *do* hold her responsible because there is more to it: Joliet took a loaded gun, with the safety off, downstairs with her, aimed it at a human form, and pulled the trigger. Panic is not a brute causal force in this story, but a crucial explanatory factor that helps make sense of the rest; yet it is the rest—which includes any number of semantic connections—that explains the relation between the panic and the trigger-pulling. That is: Sher's account is straightforwardly in line with attributionism, or else it is meaningless. But I said that this is only if we take his

[37] Sher notes that he does not mean to take any "position on the question of whether the… attitudes that account for the agent's failure to recognize that he is acting wrongly must themselves have been produced in an appropriate way." (Sher 2006 298)

argument at face value; another way to take it is as suggesting that control may not require rational—and thus semantic—connections at all. And to this point I will return.

C. Defending and Redefining Attributionism

Perhaps the strongest defense of volitionism against the attributionist menace in recent years has been laid out in a series of papers by Neil Levy. In laying out his arguments, I am indirectly presenting the upshot of this chapter: volitionism is right on its crucial point, but this requires a reconfiguration, not a rejection of attributionism. If we are to provide a theory of responsibility that is true to the phenomenon, we will have to combine the two, and this will be my project for the rest of this chapter and the following one. Levy himself has suggested that much of the disagreement may be verbal: given the attributionist distinction between blaming (and/or punishing) and holding responsible, the volitionist account may aim at the former but not the latter, which may perhaps be titled, in Watson's terminology, areatic criticism. (Levy and McKenna 2009 118) But I think attempts to bring the two camps together must be more involved. First, I have attempted to raise the stakes by arguing that the Basic Argument requires us to accept a roughly attributionist view in order to make sense of responsibility not just for attitudes and omissions, but for any action whatsoever, insofar as actions are themselves the products of attitudes, values, and patterns of thought and inference that we do not have control over (again, in the volitionist sense). Second, we will need to redefine notions such as control and choice in order to bring the two accounts together.

Levy's arguments can be boiled down to the following:

(1) Consciousness is necessary for responsibility.
(2) Attributionism fails to account for the conceptual distinction between bad and blameworthy agents.
(3) Attributionism fails to consider epistemic conditions on responsibility.
(4) Attributionism presupposes volitionism.

First, Levy argues that, on a common and plausible view about the function of consciousness, it is a necessary condition for responsibility; in fact, going a step further, Levy argues that attributability itself presupposes consciousness. "The relevant control problem arises, recall, because we do not exercise control over anything of which we are unaware" (Levy 2008 216); thus, in proposing that the agent's failure to recognize his wrongdoing can be explained by reference to his other (unconscious) attitudes, Sher does not restore control; he merely shows that agents lack it. But why should we accept the premise here? Why should we think that we do not exercise control over anything of which we are unaware? Levy begins by pointing out that, "in the absence of consciousness we are at the mercy of automatic responses we are not responsible for acquiring, that we may consciously reject and which we may even have worked hard to eradicate. It is therefore unfair to hold us responsible for actions which reflect such responses when we cannot control them." (2008 219) But this is a mix of (1) the clearly true (that in the absence of consciousness, automatic processes rule), (2) the question begging (the claim that we are not responsible for acquiring the processes; that we cannot control them), and (3) the normative (that it is *unfair* to hold us responsible). Given the attributionist willingness to hold responsible for automatic processes, this is not yet a response.

What is needed is a deeper explanation of why consciousness matters to responsibility. Levy provides this by drawing on the notion of consciousness as a "global

workspace," which "allows all the mechanisms constitutive of the agent, personal and subpersonal, conscious and unconscious, to contribute to the process of decision-making. Hence conscious deliberation is properly reflective of the entire person, including her consciously endorsed values." (2008 220) Consciousness, in other words, brings together the various attitudes and mechanisms of the person, allowing for a CTD to emerge. Levy does not claim that consciousness itself makes decisions; rather, decisions are the products of our subpersonal mechanisms; but since consciousness brings these together, the resulting action "will be controlled by *us*, in the fullest sense; by our real selves, for these mechanisms are us." (2007 242) A CTD, then, is representative of the person's values and commitments in a way that automatic processes are not, precisely because consciousness brings those processes together with values and commitments out in the open so they can be explicitly compared and contrasted. The interesting conclusion here is that a CTD is therefore reflective of the person, of the real self, which suggests that only CTDs are properly attributable to agents. Attributionists, in other words, should embrace consciousness.

This argument is hardly conclusive, however. For one thing, it is clear that unconscious (in the sense Levy, Sher, et al use) processes can be rational; and they can be fairly global, as well. Consider the "dawning" examples raised by Arpaly: if it *dawns* on you that you are in love with a friend, that you live under a corrupt regime, that Jews are not servants of the devil, the conclusion can be a fully rational one, and one that clearly draws on many of your different processes and values. Anyone involved in problem solving also knows that it is possible to solve problems in one's sleep, sometimes even when the problems are highly complex and require diverse mechanisms and values for

their solution: the solution simply dawns on you, in such cases, after a period of *not* consciously thinking about the problem. So if the mark of being representative of the self as a whole is that a decision involves bringing together different values, attitudes, and processes, there is no obvious reason why only consciousness should be capable of accomplishing this task. Nor is consciousness infallible; as Levy writes, deliberation will "greatly increase the likelihood that the resulting action reflects our real selves"; "conscious deliberation—typically—greatly improves the quality of the decisions the subpersonal mechanisms ultimately cause." (2007 241) There is nothing objectionable here from an attributionist standpoint, however: consciousness *improves* the quality and makes it *more likely* that the decision will reflect our real selves. Arpaly, we should note, agrees fully: "Another central property of deliberation and reflection is that they *focus your concentration*, allowing you to pull together mental resources from many different corners of your psyche", and this "makes it likely that more and more of your relevant beliefs will become salient to you, increasing the chances of a satisfactory solution." So CTD is extremely useful; "the ability to deliberate and reflect helps make us, as humans, be rational much more often." (Arpaly 2003 64-65)

Consciousness is very useful in helping us reach conclusions that are rational, and that better reflect our values, and this should be common ground. But it does not follow that *only* consciousness can do this; nor does it show that consciousness necessarily *will*. The same tasks can be performed without CTD, and sometimes CTD yields inferior results thanks to a distortion operating in the agent's decision-making (self-deception, wish-fulfillment, or simply a failure to consider some evidence since, of course, consciousness does not guarantee that *all* relevant evidence will become salient). If so,

then consciousness is neither necessary nor sufficient for reaching conclusions that reflect our deep, real, or whole selves, though it typically increases the chances of our doing so. Levy contends that decisions reached through automatic actions and processes are not attributable to an agent; "it might be *reflective* of the agent…but it is only by chance that it is so. The bad agent did not have the opportunity to think twice, and therefore cannot be blamed for their action." (2007 242) But consciousness, or "thinking twice," does not eliminate the element of chance; sometimes is exaggerates that element by introducing distortions. To suggest that consciousness is valuable in reaching decisions that are representative of an agent is perfectly correct. But the claim that consciousness is a pre-condition for attributability presupposes that CTD is privileged vis-à-vis the other attitudes and processes of the psyche. And this is plainly question begging.

But Levy's argument clearly gets at something important: to be reflective of the agent, attitudes or actions need a focus. That focus may not be CTD, but it does not follow that we can treat the whole self as a giant lens, magically focusing a beam of light in which the agent's true values, rational decisions, or—more importantly—attitudes or actions subject to responsibility attribution are reflected. What we will need is an ownership condition as a source of responsible agency.

Let me now take up the next two, interrelated, objections. Levy argues (1) that there is a prima facie plausible distinction between bad and blameworthy agents: it is one thing to describe an agent as bad because he is a psychopath, or because he lacks epistemic access to the moral norms he is violating (2). In response to the first argument, Smith (2008) has argued that we should be loathe to treat agents as bad rather than blameworthy, though she has not ruled out our doing so. We should avoid it, however,

because to treat agents as merely bad is to treat them as no longer subject to our reactive attitudes; that is, it is to no longer regard them as persons. But in response to at least *some* of the cases Levy cites, it seems like this is exactly what we should do. For example, in discussing Scanlon's attributionist view, Levy notes the problematic case of the psychopath who cannot see that the fact that an action harms another person counts as a reason against it. On Scanlon's view, this fact alone demonstrates that the agent is blameworthy—his disregard for my well-being is the expression of a value, and this value is an evil one. And Levy is right, I think, to note that this is a deeply counterintuitive claim. A similar issue crops up in the case of Phineas Gage, whose objectionable behavior is the result of a spike through the brain. Arpaly's response to Gage is that, regardless of what made him this way, he is blameworthy now. I think we can reject these ways of addressing the cases, however, without rejection attributionism as such.

There is a difference between a killer who is *not* a psychopath and a killer who is. One can understand that causing suffering to others is morally objectionable—he grasps the reason—but fails to respond to it because, for example, he values the enjoyment he gets from the suffering, and this valuation runs deep. We may even suppose that, because the value of enjoying suffering runs so deep, he is completely oblivious to the reason not to cause harm. Here we can explain the agent's responsibility in attributionist terms by running a counterfactual scenario: were the agent not so attached, or attached at all, to the pleasure he derives from the suffering of others, he *would* respond to the reason not to cause that suffering. But the case of the genuine psychopath may well be different: perhaps he does not see the reason *at all*, and this failure to see the reason is not a result

of his other attitudes at all. If this is a reasonable way to distinguish between the psycho killer and the run-of-the-mill sadist, then the correct response should be this: the psychopath really *ought* to be excluded from the persons club, while the sadist is fully blameworthy. Making this distinction may require a normative conception of personhood, one that excludes the psychopath because he fails to fulfill it. (I will address this point later on.) In this case we can make a distinction between the bad and the blameworthy, but we make it by appealing to a normative conception of personhood, not by taking up a volitionist stance.

What about Gage? A spike through the head, we might assume, interferes with the ability to respond appropriately to reasons: if it didn't, Gage's personality change would be hard to explain. But if Gage cannot reason properly, then it is difficult to claim that his actions and attitudes stand in the requisite rational connection to his underlying judgments and the rest of his character. In fact, the case is just the opposite: Gage's behavior is not attributable to him, because it is not an expression of his rational activity as an agent; his rational activity has been shot to hell. If we could determine just which rational connections have been disrupted, perhaps we could decide to what extent, and in what contexts, to treat Gage as a person and in which ones not to. But treating him as a full-fledged person would make no sense, since his agency is clearly not responsive to reasons in anything like the normal way. The claims so far suggest a deviation from some of the specific judgments attributionists have made, but I do not think a rejection of attributionism is thereby warranted. We will, however, need to bring in some kind of historical condition. One reason attributionists have trouble distinguishing bad from blameworthy agents, after all, is that they leave history entirely out of our assessments of

responsibility, leading to an impoverished, one-dimensional account. (Stroud 2007) Just what sort of historical account we will need to add to attributionism is a point I will return to.

Let us turn to Levy's third objection: there are strict epistemic conditions on responsibility, such that "ignorance of a moral concern excuses someone of responsibility for failing to consider it" (2005 9), and attributionists fail to account for them. To some extent, we have already addressed this point in connection with (2): agents who, counterfactually, would lack access to the relevant epistemic standards whatever their other character traits will simply have to be treated as non-persons in an important sense. But Levy has a few other examples, ones that pertain not only to psychopaths but to the rest of us as well. He gives us two examples of how lack of access to epistemic standards could serve to excuse one from responsibility (this is connected to (2), since someone who violates norms to which he lacks epistemic access can be considered bad but not blameworthy). First, suppose "that plants can be harmed, and that this harm is a moral reason against killing or treading on them. In that case, many of us are causally responsible for a great many moral harms. Are we morally responsible for them?" (2005 9) The answer Levy insists on, of course, is negative. Second, "many people have the intuition that Aristotle was not blameworthy for keeping slaves or for his sexism; that his actions and attitudes were wrong, but not blameworthy." (Levy and McKenna 2009 117) These are supposedly cases in which agents have "faulty" norms, and as a result cannot be held responsible for violating them.

But the cases are not clear. The plant example, in particular, is not one on which our intuitions are likely to be good either way, since it is hard to see what it means to say

that causing harm to plants is immoral in a way we are not aware of. Does it mean, for example, that plants suffer in ways we are not currently familiar with but that, if we only knew them, we would be likely to recognize as binding on us? In this case—if it is even coherent—we are clearly not blameworthy on the attributionist view, since ignorance *is* an excusing factor. But perhaps—and this is the only way I can interpret this—at some point in the future biocentrism will win out as the dominant account of moral considerability, so that common sense will dictate that harming plants is morally wrong. In that case, the attributionists will indeed have a harder time than the volitionists; they may even have to admit that we are blameworthy for harming plants today. But it is not obvious that they would be wrong, and it is even less obvious that they would be wrong in the eyes of the people of this imaginary future. To see this, let us turn to Aristotle.

How seriously are we to take the intuition, held by "many people", that Aristotle was not blameworthy? After all, most of my students (and not just *my* students) have the intuition that—in Aristotle's time and culture—slavery is not wrong; in which case, of course, Aristotle was not blameworthy.[38] That intuition, as virtually all philosophers would agree, is a bad one, because it fails to withstand scrutiny. How much better does the intuition that Aristotle was not blameworthy hold up? To keep this question from being rhetorical, we can look at Arpaly's account of the conscientious Nazi and similar figures. Are these people blameworthy? Her answer is: probably yes, but it depends. Imagine, first, the German whose encounters with Jews have always been negative ones, such that anyone in their position would conclude that Jews are greedy and a threat to German society. Such a person genuinely lacks any reason to accept the opposite

[38] A few weeks ago, many of my students insisted that there was nothing wrong with the Romans feeding Christians to lions—after all, this is what entertained them!

conclusion, and to that extent he is not blameworthy. But this is not the case with most Germans. A German who was aware that Jews were people like himself, but focused on their negative features and reinforced these with what he heard from the propaganda campaign is hardly free of blame: his willingness to believe the worst is not based on ignorance, but is guided by an underlying antisemitism. So what of Aristotle? Again, it depends. To the extent that he had ample opportunity to recognize that slaves are conditioned by slavery into a certain mode of behavior, and that they are otherwise no different than him, and to the extent that this failure was itself conditioned by some desire to believe in natural inequality, there is good reason to wonder why anyone would—after consideration—hold Aristotle not to be blameworthy. Nor is it exactly clear how volitionism would help: if Aristotle genuinely had no way of discerning that slaves were just as much human beings as citizens, volitionists and attributionists are in the same boat in that neither has any reason to hold him blameworthy. But if—as seems more plausible—he had access to the relevant reasons but ignored them as a result of his other commitments, volitionism does not seem to help.

What about (4), the argument that attributionism covertly relies on volitionism? First, Levy rejects the attributionist claim "that our judgment-sensitive attitudes are in principle within our control." (2005 10) On his view, it makes sense to say that "I am responsible for my attitudes if I have genuinely been (relevant) active with regard to them; if I have chosen them." But this is not what attributionists claim: they insist, as we have seen, that we are responsible for expressions of our rational activity even when we cannot change them "at will". While recognizing that *actual* control makes a difference, Levy sees no reason "for thinking that *in principle* control matters at all." (2005 10) The

attributionists have set up the problem in such a way, however, that it is difficult to make a response that is not question-begging. Levy thus provides two analogies. First, "I don't have a kind of ersatz control over my car if the steering wheel falls off; the fact that cars are *in principle* controllable does not alter my lack of control in that particular circumstance." (2005 10) Second, responding to the attributionist distinction between things that are and things that are not judgment-sensitive (e.g., attitudes vs. height), Levy imagines the scenario that we are one of very few intelligent species whose height is not judgment-sensitive. "Do we thereby become responsible for our height? Martian attributionists will claim that we are: Since height belongs to the class of things that are judgment-sensitive, *Homo sapiens*' actual inability to control their height does not alter their responsibility for it." (2005 10) The argument here is that it only makes sense to hold us responsible for attitudes that we can change at will, that we have actual control over. And this is a volitionist condition; an attributionist notion of "in principle" control seems to rely on the intuitive idea that actual control is what really matters.

But are the analogies convincing? Again, it matters what these are supposed to be analogies for. A car is, of course, in principle controllable, but a car without a steering wheel is not controllable even in principle. So, once again, we need to distinguish between the psychopath, who lacks control even in principle, and Aristotle, at least on my reading, who has control in principle but fails to exercise it. As for the Martians: they are wrong, and wrong in obvious ways. Perhaps *Martian* height is judgment sensitive, but human height is not. If they think that height—regardless of species—is judgment-sensitive, they are mistaken about the facts, for in this case we lack even in principle control. But we do not lack in principle control over our judgments and attitudes in the

same way unless we have been incapacitated by an injury or illness that blocks our rationality. If any attributionists claim that *all* humans have in principle control over *all* their attitudes, they are wrong. But this is a strike against the attributionists in question, not against attributionism, and it is a strike against them precisely because they fail to take the attributionist condition of rational activity seriously enough.

Finally, Levy takes issue with the attributionist view of justification. We can indeed ask agents to justify their judgment-sensitive attitudes, but justification is forward-looking. In discovering a flaw in you, I can bring the flaw to your attention by asking you to justify it (and, perhaps, by pointing out the ways in which it is a flaw). Having done this, I help you to satisfy the epistemic condition on responsibility since, now that your flaw has been brought to your attention, you are in a position to do something about it. But it follows from this only that you become responsible for your attitude after I have demanded a justification for it; "you are responsible for your attitude because the volitionist, and not the attributionist, conditions upon responsibility are satisfied, and, second, it hardly follows from the fact that you are *now* responsible that you were responsible all along." (2005 11) This is a strong intuitive point, but of course the attributionist does not claim that the appropriateness of a demand for justification is *the reason* why we are responsible for our attitudes: rather, the appropriateness of the demand is a sign that the thing in question is one for which we are responsible. And it can serve as such a sign because the thing in question is an expression of our rational activity.

If attributionism is mistaken, then, it is so not because it misunderstands the role of justification, but because it drops the control requirement rather than revising it. Here we can make a start toward remedying that, for the attributionist can ask: If I am not

responsible for my attitude, how, exactly, does my being asked to justify it *make* me responsible? If the attitude is an expression of my practical agency, or my rational activity, then—provided my rationality is working correctly and has access to the relevant facts—it seems nothing more is needed. The antisemite who has no evidence that Jews are anything but vicious schemers *can* become responsible if we introduce him to Jews who clearly aren't; he does indeed become responsible, but only because he was lacking evidence that was required for him to, counterfactually, have been responsible in the first place. The antisemite who ignores some evidence and sticks to other evidence as a result of his attitude toward Jews, on the other hand, will not be likely to change his mind when asked for a justification. And if his attitude *before* the demand for justification was not blameworthy, it is unclear why his attitude after will be any more so. No doubt *I* might now hold him responsible, having demanded justification from him; but this is an epistemic requirement on my holding another responsible, not on his responsibility itself. When a justification is demanded of me, I may either ignore it or change my mind (perhaps over a long period of time, as such things frequently go), if you convince me that I have been overlooking something, say. If I ignore it, it is not clear that anything new has been added; if you convince me and I change my mind, this shows only that I was *capable* of changing my mind all along—that I was already in control.

Smith, as we have already seen in sketching the reply she might give to McKenna, has something very much like a control condition built into rational activity as such. We are responsible for a state, on her account, if it is *in principle* open to revision. This in principle openness amounts, I suggest, to in principle control. Of course, as I have suggested in response to Levy's criticism of this point, the state must be genuinely open

to revision: the rationality of the agent must be intact enough that it could, counterfactually, change the attitude; and the agent must have access to the relevant facts or reasons. The control involved in having, maintaining, or changing the judgment or attitude in question need not be conscious, but that does not mean it is not control. Smith notes that, "if one thinks that 'choices' can, like judgments, be inexplicit, unconscious, and attributed to a person simply in virtue of her responses, then my disagreement with the volitional view would turn out to be much less significant." (Smith 2005 256) The suggestion, then, is that we could reconcile attributionism with volitionism if only we could provide accounts of "control" and "choice" that do not demand conscious awareness.

In his thorough review of Arpaly's book, Robert Pippin seems to be making a similar point. Arpaly, he argues, overstates her case; if her point is largely that we can and do deliberate, respond to reasons, and make decisions without explicitly sitting down to think, being able to fully describe what we are doing, or even having a grasp on the principles on which we may be acting, this point is not necessarily a challenge to standard moral psychology. Oliver Single may not have made a fully aware decision to turn state's evidence against his father's firm; but his action was not like an involuntary spasm, either. "Single's 'gradual disaffection' is an expression of a change of allegiance *he* is coming to effect rather than merely undergo, and I see no reason why we should not call that a matter of ongoing everyday deliberation over a long period of time." (Pippin 2007 293-294) The suggestion is that a non-explicit process of which the agent is largely unaware may well be called "deliberation" without much change in the ordinary meaning of the term. Can we not do the same for "control" and "choice"?

This sort of redefining strategy is at the heart of Sher's argument and his attempt to lay out a new, non-volitionist notion of control. But we miss a crucial point here if we focus on the question of whether Sher manages to meet the right epistemic conditions for control, since control is not the only element in question. After pointing to the traditional idea that "an agent's control extends no further than the searchlight of his conscious awareness," Sher notes that "underlying this assumption was a certain familiar conception of the controlling agent himself—one which takes him to be simply a center of consciousness and will." (2006 296) Though I do not find Sher's account of control especially satisfying—at least compared to the reconfiguration of Arpaly's view, sketched above—this is a crucial claim. The phrase "the agent's control" has two terms: an agent, and his control. But we cannot focus only on the control aspect, since our notion of agency is obviously relevant here: if we compare the phrase "the agent's control" with the phrase "the sponge's control," we clearly mean different things by "control" if both phrases are to be referring to something coherent, and very different epistemic conditions will apply. That is: the conditions for control, and the definition of the term itself, will depend heavily on how we understand the agent. The central difficulty with relying on intuitions about the relation between control and consciousness is that they import many underlying notions about the agent, the person, or the self, including the Cartesian view that the self is co-extensive with consciousness. I am not claiming that objections to attributionism rest on the Cartesian view as a premise; only that they may smuggle the view in through the backdoor. Someone who rejects the Cartesian picture may still find it active in his intuitions. And, we might wonder, if we

manage to remove the picture from our intuitions, would the need to establish a connection between control and consciousness still be as obvious or as pressing?

I have been arguing that attributionism is, at least partly, the right account of moral responsibility, and that it is necessary if we are to offer a response to the Basic Argument. But my point here is not that volitionism is off the mark; its mistake is only to assume that the "control" and "choice" of the kind needed for moral responsibility must be understood in terms of CTD. As I will argue in the following chapter, to provide a complete account, attributionism needs to be augmented with redefined and clarified notions of these volitionist terms, without which the link between the agent and the things for which she is responsible remains unclear. Furthermore, attributionism requires a stronger account of what it means for an attitude or action to be the agent's *own*. Finally, both of these projects will require adding a temporal dimension to the attributionist account, or we run the risk of holding the wrong agent responsible—a problem of *blaming* the *bad* agent. As Levy notes, it is perfectly possible, not to mention common, to be a globally responsible agent while having islands of attitudes for which one is not responsible. (Levy and McKenna 2009 118) Without ownership, control, choice, and historical conditions, we will be unable to make sense of this division within the self.

5 Control and History

A. Control

In the last chapter I suggested that there are two ways to understand Sher's disagreement with attributionism. The first—what I called taking the criticism at face value—is to see Sher as arguing that the connections between an agent's beliefs or attitudes and her action need not involve a semantic component in order to allow for attributions of responsibility; that connection need only be an appropriately causal one. I attempted to cast doubt on the idea that we can make any sense of the notion of appropriateness involved without bringing in some semantic component. But I suggested also that there might be another way to take Sher's criticism: as arguing that control may not require rational (or semantic) connections at all. We can take up this point by noting something peculiar about many of the cases that anti-volitionists (Sher included) marshal in defense of their claim: whether involving miscalculations while driving, forgetting about something while involved in work or conversation, or tossing one's niece in the air, they are textbook examples of what Hubert Dreyfus has called "absorbed coping." They are cases, in other words, of situations where agents act without any explicit thinking, usually because the activity they are performing is one they have mastered to such an extent that thought is unnecessary and could, in fact, interfere with the performance of the

action.[39] Anti-volitionists, in general, want to argue that such cases, despite the lack of CTD, nevertheless allow attribution of responsibility.

The disagreement I am here suggesting between Sher and the others is over whether a lack of CTD in absorbed coping involves also a lack of rationality. The standard attributionist case rests on hanging responsibility on the rationality implicit in absorbed coping; as I have been suggesting, this rationality moreover allows us to attribute agential control, so that we need not separate responsibility from control. But the position becomes more tenuous if we shift to Sher's view and leave rationality out of absorbed coping altogether. In a recent attack on John McDowell's work, Dreyfus has defended just such a position. My aim here will be to argue that Dreyfus (and Sher, on this understanding of his argument) cannot be right; that cases of absorbed coping involve rationality capable of grounding agential control, and that, furthermore, this control is precisely what allows for responsibility attribution.[40] Dreyfus's position in this debate is far more extreme than his previous work, and I do not claim to be accurately conveying his views (or, for that matter, those of McDowell). The point is only to suggest, on the basis of what is at stake in the debate, a rough model of agency that allows for rationality in the complete absence of CTD.

In *Mind and World*, McDowell argued that contemporary epistemology is caught in an oscillation. On the one hand, we face the Myth of the Given, the idea that passive experience can have only a causal effect on our beliefs. The difficulty with this view, as

[39] Not all the cases described by anti-volitionists are examples of absorbed coping. Nevertheless, these cases, which leave out any possibility of explicitly made blameworthy judgments, pose the greatest challenge to volitionist accounts.

[40] Of course the presence of rationality within absorbed coping is not sufficient for responsibility. After all, it is not even sufficient for agential control. The point here will be only that action—even absorbed action—involves rationality, and is thus subject to being carried out by an agent.

amply demonstrated by Wilfrid Sellars, is that it makes it impossible to explain how experience can provide any *justification* for beliefs. The other side of the oscillation results from the temptation to recoil from the Myth by insisting that experience does not justify our beliefs—only other beliefs can do that, as Davidson argued. Both approaches are unsatisfactory, and McDowell argues that the way out of the dilemma is to recognize "that even though experience is passive, it draws into operation capacities that genuinely belong to spontaneity." (McDowell 1996 13) In other words, the difficulty is that our experience clearly plays some role in justifying our beliefs, and we cannot account for this fact without taking experience to already have a rational structure. On this picture, then, "our perceptual relation to the world is conceptual all the way out to the world's impacts on our receptive capacities."(McDowell 2007b 338) To put it another way, "our perceptual experience is permeated with rationality." (2007 339)

Just as McDowell's view involves an account of perception, it also comes bundled with an account of rational agency. To perform a rational action, he points out, one must be exercising one's conceptual capacities in two ways. First, one must have a conceptual experience of something in the world that solicits one to action and, second, one's action itself must be the actualization of a conceptual capacity. That an agent act rationally in this sense, however, requires neither prior deliberation nor the explicit grasping of something in the world as a reason for acting. Rather, for some action on an inclination (e.g., an inclination to flee from danger) to be action for a reason, "we would need to be considering a subject who can step back from an inclination to flee… and raise the question whether she *should* be so inclined—whether the apparent danger is, here and now, a sufficient reason for fleeing."(McDowell 2009 128) But McDowell stresses that

acting for a reason does not involve stepping back and explicitly considering it; "acting for a reason, which one is responding to as such, does not require that one reflects about whether some consideration is a sufficient rational warrant for something it seems to recommend. It is enough that one could." (2009 129) This means that for some inclination I to be a reason for action R, and for an agent to act *on* that reason, what is needed is only that the agent be capable of taking I to be a reason R, i.e., capable of using I in practical deliberation, though the deliberation need not and may never occur.

To illustrate the point, McDowell imagines the case of a hiker on a marked trail,

> who at a crossing of paths goes to the right in response to a signpost pointing that way. It would be absurd to say that for going to the right to be a rational response to the signpost, it must issue from the subject's making an explicit determination that the way the signpost points gives her a reason for going to the right. What matters is just that she acts as she does because (this is a reason-introducing 'because') the signpost points to the right. (This explanation competes with, for instance, supposing she goes to the right at random, without noticing the signpost, or noticing it but not understanding it.) What shows that she goes to the right in rational response to the way the signpost points might be just that she can afterwards answer the question why she went to the right—a request for her reason for doing that—by saying 'There was a signpost going to the right'. She need not have adverted to that reason and decided on that basis to go to the right. (McDowell 2009 129)

We should not, I think, take McDowell's account to mean that the ability to retrospectively reconstruct one's reasons for action is a *necessary* condition for acting for a reason. It would, after all, not even be a necessary condition if acting for a reason did require explicit deliberation. Compare, for example, the case of two hikers, A and B. A deliberates whether to follow the signpost and, after deliberation, turns right. B fails to register the signpost in any way, and turns right entirely at random. Surely the difference between A and B cannot turn on whether or not A remembers the deliberation: A might, for example, have an awful memory, so that she is incapable of reconstructing even her

explicit deliberative processes. “Why did you turn right back there?” “I turned right back there?” The difference between someone who acts for a reason deliberately and someone who acts for a reason without deliberation, then, is not a matter of one acting rationally and the other acting irrationally, but merely a matter of whether the reason is taken up into practical deliberation. The difference between someone who acts for a reason and someone who does not, on the other hand, will be much harder to specify.

This is especially true given McDowell’s various specifications of his notion of conceptuality. First off, a concept—as we have seen—is just something that can be taken up into explicit reasoning. But in response to Dreyfus’s insistence that our embodied coping with the world does not and cannot make use of abstract, general principles, McDowell notes that concepts can be situation-specific. They might, for example, take the form of demonstratives, so that a shade of color for which agents have no word might simply be referred to as a “this,” while a bodily movement the agent is adept at performing may be intended “under specifications like ‘whatever is needed to throw efficiently to first base.’” (McDowell 2007a 368) Thus, McDowell can accommodate Dreyfus’s point—that embodied coping leaves no room for abstract thinking, and that in fact abstract thinking about how to act interferes with the ability to act efficiently—without giving up on the idea that rationality may well be present within the embodied coping itself. Moreover, it is not necessary—for content to be “conceptual” in McDowell’s sense—for the agent to have a word for the concept or even to ever bring it into her linguistic repertoire. Thus, McDowell distinguishes between experience that “is embraced by conceptual capacities… that we already had before we enjoyed the experience,” and experience that can be isolated and articulated by “annexing bits of

language to" it, even though "some of the content of a typically rich world-disclosing experience never makes its way into constituting part of the content of our repertoire of conceptual capacities." (McDowell 2007b 347) So while obviously not *all* of our experience—not even most of it—is ever articulated, "*all* its content is present in a *form* in which… it is suitable to constitute contents of conceptual capacities". (2007b 347) McDowell thus introduces a category of what I will call pre-conceptual experience; that is, experience that has not yet been assimilated under a concept (and perhaps may never be so assimilated), and yet which has, by virtue of its form, the potential to be fully conceptually articulated.

Dreyfus, on the other hand, argues that McDowell has fallen into "the myth of the mental," namely, the myth that all our capacities are permeated with mentality, "declaring that human experience is upper stories all the way down". (Dreyfus 2005 47) This myth, Dreyfus argues, effectively ignores the embodied coping going on at the lower stories, and thus overlooks the background necessary for any rational thinking to occur in the first place; the world draws us to action with its solicitations and without our explicit deliberation, and "these solicitations have a systematic order that… works in the background to make rationality possible, but the system of solicitations is not itself rational." (Dreyfus 2007c 358) Embodied coping, which involves action that is expert and thus does not require deliberation, consists not of openness to a conceptually articulated world, but of immediate bodily responses to perceptual deliverances. This allows us, Dreyfus thinks, to make sense of the idea that "animals, prelinguistic infants, and everyday experts like us" all share the same space, which only we can step back from, but "when a master has to deliberate in chess or in any skill domain, it's because

there has been some sort of disturbance that has disrupted her intuitive response." (Dreyfus 2005 57) The intuitive idea, then, is that in interacting with the world we seem to share certain capacities with the other animals, and that we—expert copers that we are—only transcend those shared capacities when something has gone wrong with our coping, or in order to learn a new skill.

On Dreyfus's view, this does leave phenomenology with an unresolved challenge: just how do we manage to move from the non-conceptual lower story to a conceptual upper one? By contrast, we might characterize McDowell's account as responding precisely to this problem: the question, once asked, cannot be answered. Unless we recognize that our perception is conceptual all the way down, we will never be able to get back up again. McDowell thus stresses that, although we do, in some sense, share capacities with the other animals, the deliverances of our perception are always already of a form suitable for rationality—to continue with the upper stories analogy, we might present McDowell's account thus: humans and animals might have the same perceptual matter, but in humans that matter *lives in* the upper stories. Now if we were to accept McDowell's position, it would indeed help resolve the problem Dreyfus leaves unsolved. The issue, however, is whether this model makes sense. Dreyfus seems to take it as an ad hoc mechanism, which falsifies our phenomenological experience of agency. Dreyfus's early arguments against McDowell—which rely on the notion that any explicit rational thought necessarily interferes with our ability to perform engaged bodily tasks, and that our engaged coping responds directly to solicitations from the world rather than relying on any general rules—are easily disposed of once McDowell's view of concepts as both

situation-specific and not necessarily explicit is laid out. Dreyfus's remaining objections, then, come down to the following three:

(1) Rationality is not part of our phenomenology; nowhere in our absorbed coping with the world do we encounter our conceptual capacities at work. The "conclusion [that our coping is permeated with rationality] is supposed to follow from the fact that if one has a *capacity*—in this case the capacity to use situation-specific concepts—this capacity must be '*operative*', as McDowell puts it, in all situations whether or not I am aware of *exercising* it." And this, we are told, is a "category mistake": "Capacities are exercised on occasion, but that does not allow one to conclude that, even when they are not exercised, they are, nonetheless, 'operative' and thus pervade all our activities." (Dreyfus 2007b 372) Although we can step back from our engaged experience in the world and contemplate its affordances (e.g., doors afford going through; telephones afford dialing), we cannot respond to its solicitations as such if we are thinking about them; McDowell's view of our openness to the world, "while true to our experience of affordances as facts, flies in the face of the phenomenon of *solicitations*… there is no place in the phenomenology of highly skillful action for conceptual mindedness." (Dreyfus 2007c 361)

(2) Experts or masters do not follow rules in their expert activity; they act on immediate perception, and it is difficult to see how reasons can play any role in such action, especially when we notice that "when an expert is forced to give the *reasons* that led to his action, his account will necessarily be a retroactive *rationalization* that shows as best that the expert can retrieve from memory the general principles and tactical rules he once followed as a competent performer." (Dreyfus 2005 54)

(3) Following on this point, we can note that if the deliverances of sensibility cannot be fully articulated, or can be articulated only in extremely wide-range demonstrative concepts, it becomes unclear what sense there is in saying that an agent is acting for a reason when she follows a particular solicitation. Something about the world draws her to act, but what it is sometimes cannot be fully articulated. Using one of his favorite examples, the Grandmaster, Dreyfus notes that "pointing to the specific pieces on the specific squares on the board as *that position* doesn't capture what it is about that position that draws the Grandmasters to make that move." (2007a 105) First off, it is not merely the position that draws the Grandmaster to move: the tempo, the opponent's style, and myriad other factors contribute as well. Ultimately, it may be that the Grandmaster's only explanation for his move would be that this was the move that made sense, or that he felt like making this move, "but such a response would be too situation-specific to count as a reason." (2007a 107) The challenge, then, is that there is nothing in what motivates the Grandmaster that it would make sense to call a reason; by extension, "something similar happens to each of us when any activity

> from taking a walk, to being absorbed in a conversation, to giving a lecture is going really well." (2007b 373)

Responding to these challenges, as I have indicated, is important to establishing that control can be present in agency in the absence of CTD. Before moving on to that discussion, I want to motivate it a bit further by pointing out that Dreyfus's account cannot make coherent sense of our responsibility for the agency undertaken in absorbed coping. After pointing out that absorbed coping goes on without any sense of a subject or of reasons for acting, Dreyfus notes that "of course the coping going on *is* mine in the sense that the coping can be interrupted at any moment by a transformation that results in an experience of stepping back from the flow of current coping. I then retroactively attach an 'I think' to the coping and take responsibility for my actions," even though within the experience itself, what is encountered are only solicitations drawing out responses. (2007c 356)[41] This account is already odd, given Dreyfus's constant insistence that our essential feature, or most pervasive kind of freedom, is not the ability to step back, but rather to become absorbed in our activity, since—as it seems—an action is only *mine* by virtue of that stepping back. But, more problematically, the references to responsibility here strike me as incoherent.

The idea seems to be that I can be responsible for my absorbed agency because I—as a thinking subject and not simply a coper—can always jump into my activity and stop it if something is not going well. Dreyfus gives the analogy of an airport radio beacon, which only gives a warning signal if the plane goes off course; but when the pilot is followings its beam, "the silence that accompanies being on course doesn't mean the beacon isn't continuing to guide the plane. Likewise, in the case of perception, the

[41] Dreyfus repeats the point in his next reply to McDowell: "My coping is mine in that I can break off doing it, and for that reason I can take responsibility for it." (2007b 375)

absence of tension doesn't mean the body isn't being constantly guided by the solicitations. On the contrary, it means that, given past experience in this familiar domain, everything is going exactly the way it should." (2007c 358) But this idea is puzzling, since Dreyfus's point is that there is no monitoring going on within the experience of absorbed coping. If so, the beacon analogy raises difficulties for both Dreyfus's account of responsibility and for the coherence of his overall attempt to excise conceptuality from most of our agency. It is, first of all, unclear just how responsibility is supposed to enter into the picture. Suppose that I am absorbed in coping, without the operation of any conceptual capacities. Responsibility does not kick in unless something goes wrong—or, perhaps, I simply step back to think about what I am doing—and I *take* responsibility for how my body has been responding to solicitations. But why, exactly, should I take responsibility for something my body has been doing? *I* don't enter onto the scene as long as the absorbed coping is going smoothly; unless my body has been following *my* guidance all along, taking responsibility for it seems like an odd maneuver; at most, it would be by default—perhaps someone needs to take responsibility, and no one else is available. But if there is no rational link between myself and what my body has been doing there is no particularly good reason for me to take responsibility for its activity. Far from explaining how responsibility for absorbed coping—that is, for the vast majority of the work of our agency in the world—is possible, Dreyfus's account seems to obscure the possibility of anyone's ever being responsible for it.

In the case of the beacon, the pilot may be greeted with silence so long as the plane stays on course, but the system can work only because the beacon itself is monitoring the plane's trajectory. But now imagine, as Dreyfus would have it, that

nothing is monitoring the absorbed coping. If something goes wrong—some tension arises—then explicit thinking kicks in and our responding to solicitations is transformed by being bumped up to the upper stories of rationality. Something exceptionally strange has happened here: a rational system has kicked in, but it has—presumably—kicked in for no reason! After all, there was no rational monitoring of the body's responding to solicitations, nor was any rational faculty patiently looking for a tension to arise, since this would make the tension itself—a sign that something was going wrong—into a rational activity. Dreyfus admits that phenomenology has difficulty explaining how our non-conceptual coping can be transformed into conceptual activity. But the problem here goes deeper: since neither the coping nor the tension can be conceptual, on Dreyfus's view, any appearance on the scene of our rational capacities would necessarily be lacking in rational motivation; not only would it be unclear why we should take responsibility for our coping, but responsibility would also be absent from our rational interference—or lack thereof—with such coping. If the relation between our absorbed coping and our rationality is to make any sense, then, it seems our best option is to recognize that coping as such is already permeated by rationality, which is precisely why its malfunctioning can provide *reasons* for our reasons-responsive explicit thinking to step in when needed. And this, of course, is a McDowellian point.

We thus have a first stab at answering Dreyfus's first objection: thinking of our coping as already permeated with rationality makes sense of the idea that we are capable of rationally interfering with it, or of "stepping back" just when we need to. Nor is the idea that some monitoring experience must always be in the background of all our coping—rather, the point is that since the coping is already conceptual and thus capable

of providing reasons to our reasons-responsive deliberative apparatus, the monitoring goes on at the level of the *form* of our coping and not through some additional conscious process. We may add another observation: Dreyfus argues that rationality cannot be present in coping, since much of what goes on in our expert activity happens too quickly to involve any thought. If this is right, we have a problem. Rational thought, it would seem, can enter on the scene almost instantaneously; and it can halt whatever absorbed coping is taking place. But how can it do that if the absorbed coping is not really coping with anything conceptual at all? In other words, how can rational thought interfere with non-rational activity unless that activity is already pre-rational, i.e., of the correct form to interface with higher level cognitive abilities? Unless such an interface is in place, rational thought will not be interfering with absorbed coping at all by grasping and evaluating affordances and solicitations; at best, it will simply crowd out those affordances and bodily responses to them and replace them with acting for a reason, a mechanism of a completely different kind that will have no reference to what the agent was doing before the shift to rational thought took place. On Dreyfus's conception, what happens in a breakdown is not that our affordances become conceptually explicit; whatever does present itself to explicit thought can have nothing in common whatsoever with those affordances, since they lack the potential to be taken up into conceptual thinking. This, incidentally, is another variant of the earlier problem of how I could take responsibility for whatever I was doing in the mode of absorbed coping: whatever it is I take responsibility for, it could not be the absorbed coping itself, since *that* is not something I could attach an "I do" or "I think" to. That McDowell's approach helps to

bridge this gap suggests phenomenological evidence for that approach rather than a rejection of it.

In arguing that we have no grounds to posit any sort of conceptual activity that is not experientially present, Dreyfus has apparently reverted to the flaccid, though currently popular, view of phenomenology as description of surface-level phenomena as they are experienced *at the time* they occur.[42] It is true that, when we are engaged in absorbed coping, we are not explicitly aware of any conceptualization occurring. But we should not take such experience out of context, since something happens *after* my absorbed coping as well: I reflect on it (not to mention, as Dreyfus admits, that I can both attach an "I think" to it and take responsibility for it). And something happens before: I am aware, generally, of what I will be doing (though of course I need not have it planned out) and, in the past, have performed similar tasks *with* explicit conceptual guidance in play. Dreyfus admits this point, but he thinks that *after* one has gone through the learning phase, where one is guided by concepts, one transcends that stage, becomes an expert, and no longer needs concepts at all. But this is quite odd: if I needed concepts to play chess in the past, is it not reasonable to think that, as I've gotten better, I have lost the need to rely on keeping those concepts explicit? But how can this be evidence that they are *not* present? Perhaps the Grandmaster is no longer following *rules*, explicit or otherwise, but his perception of the game ought to be structured by the conceptual repertoire he began with, now even more finely articulated. And it may be precisely because of the finer articulation that the Grandmaster does not need to think explicitly

[42] Again, "reverted" is the operative term here, since Dreyfus's conception of phenomenology is much deeper in his earlier work.

about what move to make: the conceptual relations in play already link up with each other to produce outputs below the level of explicit thought.

Moreover, a phenomenological account should recognize that, after my absorbed coping, I often know what happened during that time. If asked why I made a certain move, I can give a reason, although I may have to think about it in order to make it explicit. No doubt I cannot explain *every* feature of my actions, but so what? The fact that I cannot describe every feature of a blade of grass I saw does not mean that I did not see something that fits under the concept "grass." Why, then, should we focus on the unthematized experience as authoritative? Absorbed coping does not, of course, thematize the experience that goes on within it—it is of the nature of absorption that only some object of it, but not the experience itself is thematized. By Dreyfus's reasoning, we would have to conclude, more or less, that absorbed experience has *no* structures or features, since none of them are explicitly the object of awareness within the experience itself. Phenomenology on this account loses all ability to provide anything other than superficial introspective reports. A correct description of coping experience is going to be misleading, precisely because it involves an attempt to describe an experience that, by definition, was not explicitly thematized at the time it occurred; reconstructing this experience is one of the main tasks of phenomenology. McDowell's argument—that if our basic perceptual experience lacked suitability for conceptualization, it is unclear how we could articulate it at all—can thus be taken up into a phenomenological analysis of absorbed coping, since presumably the phenomenon of switching from absorbed to thoughtful agency itself requires clarification.

We can now address Dreyfus's latter two objections. Absorbed copers often cannot give the reasons—or even give wrong reasons—on which they acted. Dreyfus cites the case of fighter pilots who, studies show, do not in fact make the movements they teach to beginners as a set of rules; if called on to reconstruct their flight decisions, however, they appeal to those rules. Here is a clear case of copers performing expertly but, when called on to give their reasons, giving the wrongs ones. But this example does not by itself cast doubt in the idea that there are in fact reasons for which the copers act. Though McDowell, in his discussion of the hiker turning right at a sign, does seem to suggest that retroactive reconstruction of reasons is evidence that the agent acted for a reason, he cannot be taken to be claiming that the ability to give such a reconstruction, and to give it veraciously, is a *necessary* condition of having acted for a reason. In defense of this thought, we can marshal three of the points I have raised above. First, agents are likely to forget why they acted—they may even forget *that* they had acted—especially in cases where no explicit thought went into their action. But second, precisely because the conceptual repertoire of an expert is more varied and more finely articulated—possibly to the point where the expert lacks pre-existing expressions for much of the content of her experience—accurately repeating the conceptual pathways may well be impossible. (Think of the common experience of trying to reconstruct the course of a conversation from memory!) And in such a situation, it would not be surprising if the agent simply fell back on the simple explanations provided by pre-learned rules, even though these were not in fact involved in her action. Finally, many of the concepts involved may be demonstrative ones, for example ones referring to what are basic actions for an agent. An attempt to reconstruct one's reasoning explicitly, in trying

to make a demonstrative point to a listener on whom the situation-specific demonstrative contents are lost, will result in nonsense.

But none of this shows that agents engaged in absorbed coping cannot be acting for reasons.[43] Dreyfus's insistence that—at least in many cases of absorbed coping—whatever the agent acts on is too general to be a reason seems mistaken. Let's return to the example of basic action. Say I have mastered the skill of blocking a fencing strike. A novice now asks: how did you do that? Whatever account I give the novice will, indeed, seem too general and unhelpful. I blocked it because I saw the strike coming, and that's that. And the novice can learn my reasons, but only by being taught to see them, that is, to perform the basic action himself. Dreyfus takes himself to be basing his account, via Heidegger, on Aristotle's notion of the *phronimos*, who simply perceives the right thing to do. McDowell correctly points out that, in Aristotle's account, the perception involved in *phronesis* is clearly not outside the domain of *logos*. The *phronimos* acts rationally, although only another *phronimos* can understand his reasons. In training for virtue, the initiate learns to reason correctly by first applying general rules—e.g., steer away from extremes and aim for the mean, act justly or generously—and eventually learning to grasp the "ultimate particular fact," i.e., the action required.[44] But if the *phronimos* is asked why he performed the action, he may be able to say nothing better than "this is what justice required." To the initiate, this may indeed seem too general to count as a reason; but it may nevertheless serve as the *phronimos*'s reason for action. Dreyfus

[43] We can compare the account to Arpaly's examples, illustrating that one may well be acting rationally without knowing it.

[44] This is the rendering of *Nic Ethics* 1142a25 in the Oxford World's Classics translation by David Ross (revised by Lesley Brown).

suggests that the *phronimos*, or the Grandmaster, might simply say that the action or move was just what "felt right to him," or that he "felt like doing it." But this explanation in terms of feelings and motives is obviously not intended by Aristotle to crowd out an account that makes it the deliverance of a properly trained practical reason.[45] That an agent "felt like doing it" is far more general than the explanation in terms of reasons—after all, one might feel like doing in any number of situations where there is no reason whatsoever to do it. "This looked like the right move to make" is a much better explanation than the simple "he felt like making that move," since the former at least implicitly relates the making of the move to some goal—winning the game—while the latter leaves the entire question of why the Grandmaster did what he did entirely open; perhaps he was simply tired of playing and felt like surrendering his queen.

Now the question of how exactly conceptually articulated content is connected to the pre-articulated field of experience is an incredibly difficult one, and I will not even make a gesture at addressing it here. I will merely point to three options in attempts to outline the interaction of these two spheres within the domain of responsible action. First, we might simply state that we somehow learn certain concepts, while much of our experience remains wholly non-conceptual. The conceptual then enters into our activity, and it is this and only this sort of activity that counts as responsible action. For example, a perception with some propositional content, together with the propositional contents of a desire and a belief, entails the propositional content of an intention, which content is then (paradigmatically) realized in or by the action. Second, Dreyfus's approach has the

[45] Just how motivation and reason may be combined in this way is, of course, a difficult question. For an interesting stab at this, see Korsgaard's "Acting for a Reason," reprinted in her *Constitution of Agency* (2008a). On her account, the agent simply sees the whole action—e.g., going to Chicago to visit a sick relative—as something worth doing, takes that recognition as the reason to perform the action, and is motivated by the recognition.

conceptual arise out of the non-conceptual. We can act, and act responsibly, by responding bodily to solicitations in the world; conceptual articulation has no place in this picture. Where it appears, in fact, it indicates a breakdown in action, involving a withdrawal of the agent from the world. There is a difference in kind here between non-conceptual and conceptual content, such that to represent the former as the latter is to distort its true nature. Finally, McDowell's account, by contrast, draws a distinction between the conceptual and the pre-conceptual, rather than the non-conceptual. The field of experience, on this account, though largely unarticulated, *can* be conceptually articulated precisely because experience—at least that of rational agents—is of the kind to allow articulation. When we act, we may well act smoothly on the basis of solicitations presented by the world; but these very solicitations can and must, in order to enter into our experience at all, allow for conceptual articulation and redescription.

Philosophers sometimes write as if human beings have their heads filled with rather odd metaphysical entities: propositional states. Our beliefs, desires, attitudes, and so on are or contain propositions, conceptually articulated wholes which must be accessed as such in order to allow for genuine, and certainly for responsible, action. The alternative to such a view, it sometimes seems, is the thought that instead we are simply filled with blind natural processes and proddings, irrational pulls and pushes that—should they issue in bodily movement—involve no more agency, and carry with them no more responsibility, than an uncontrollable reflex. McDowell's position allows for a compromise. Our beliefs, desires, attitudes, and so on are, largely, inchoate. They have form, but the form is potential; it is actualized when this matter of our experience bonds with our conceptual schemes, allowing for a rational articulation. Experience, in

appearing as conceptually articulated and capable of entering into rational deliberation, thus shows itself to have been conceptual all along. And it is just such a position that allows us to say that, although much—probably most—of what we feel and do is not yet conceptual, not yet articulated, and presses on us and through us without much by way of explicit and conscious rational thought, nevertheless constitutes who we are as responsible agents and not as brutes.

Something like McDowell's position, then, allows us to articulate a thought central to attributionism: that if rationality in action is essential to responsibility attribution, this need not drive us into a volitionist corner where the only action that counts as responsible is action that involves CTD. Such an account allows us to make sense of the suggestion that our attitudes (Smith) and actions (Korsgaard) *embody* our judgments (or, to put it another way, the judgments are partially constitutive of our attitudes and actions) while allowing that the judgments may be entirely outside the agent's awareness. Our evaluative judgments are not free-floating entities grounding our attitudes. They are, rather, embodied in those attitudes. This is why the claim that we can discover our judgments through self-observation (Smith 2005 252) has significant force: changing an underlying evaluative judgment is not like realizing, in light of a demonstration, that one's response to an exam question was wrong. It involves changing an entire pattern of behavior. Changing the pattern of behavior, in turn, counts as changing an evaluative attitude. And we can allow that our attitudes and patterns of behavior—the background on which we act—are far too rich to fully articulate conceptually without thereby giving up on the notion that rationality is in play even in those behaviors and attitudes that one has never consciously chosen.

On this reading, then, McDowell's work comes far closer to Heidegger than Dreyfus's supposedly Heideggerian stance. We can see this from a footnote in which Dreyfus in fact attempts to assimilate Heidegger's position to his own without having to admit his divergence from the latter:

> This disclosing function of perception we share with animals and infants. Heidegger, however, connects such understanding with our understanding of our identity... To open a *world* in Heidegger's sense requires that the affordances that matter to us and draw us in depend not merely on our needs and previous experience, as with animals, but on what matters to us given our identities, and we are capable of changing our identities and so our world. This is an important difference between human beings and animals, but since we are focusing on the role of perception in giving us a background on the basis of which we can perceive objects and justify our beliefs about them, we needn't go into it here. (Dreyfus 2005 65n.54)

In this attempt to separate perception from "our understanding of our identity" Dreyfus diverges fundamentally from a theme crucial to *Being and Time*, which is precisely that any such separation is a theoretical construct that obscures the phenomenon of being-in-the-world. Here I want to explain my understanding of Heidegger's position for future reference; I will lay out the discussion only in basic outline, as the standard features of Heidegger's views are well-known.

Key to Heidegger's account is the notion that Da-sein, his term for the kind of being of human beings, is always in a world. This world is always a referential totality, in which entities disclose themselves to us first as tools, that we use for, with, and in common with other entities like ourselves with whom we share practices of interacting with the world and understanding ourselves. I will save a discussion of this latter aspect for the next section; here I want to focus only on the way that tools—and our coping with them—become meaningful in Heidegger's account, as this is one key part of his account of human agency, and especially of absorbed agency.

In his famous tool analysis, Heidegger lays out the conditions for there being entities for us to interact with. These entities are disclosed primarily as tools: as equipment to be used for some task. But, as Heidegger stresses, tools do not disclose themselves as individual objects; "strictly speaking, there 'is' no such thing as *a* useful thing. There always belongs to the being of a useful thing as totality of useful things in which this useful thing can be what it is."[46] This totality, furthermore, is always a referential whole: a pencil appears only in reference to writing or underlining, to paper, notebooks, and books, to the drawer in which I keep it, to my need for keeping notes, to my writing. And in its own turn, the pencil has a certain materiality—it is sharp or dull, fitting or unfitting for the task relative to which it is a pencil; what makes it so suitable is, in part, a further set of references to graphite and wood, references that—obviously—never need to appear in order for the pencil to be put to use.[47] So long as the pencil remains sharp, in its place, and otherwise suitable for underlining or writing, I need not explicitly notice it *as* a pencil—I merely use it to write. When it correctly plays this role in the referential totality, the pencil appears as *Zuhandenheit*, ready-to-hand; which is to say that it does not appear as a pencil at all but rather withdraws, in Heidegger's formulation. In using things—or taking care of them—in this way, we need not explicitly be aware of them at all; the experience or seeing of things that conforms to this mode of taking care is called circumspection (*Umsicht*).

---

[46] Heidegger 1996, 68. All page references to *Being and Time* will refer to the pagination of 1953 Max Niemeyer Verlag German edition, given in the margins of both English translations. I will quote from Stambaugh's translation, though some terminology will occasionally need to be altered. From this point on, when page numbers are given without further information, they always refer to the German pagination of *Sein und Zeit*.

[47] Heidegger commentators frequently focus entirely on the social dimension of tools—that is, the norms that govern their use—as marking their user-independence. That the materiality of the tools constitutes another level of user-independence—one that, moreover, both places constraints on and is taken up in norms of use—is a point recently worked out in detail by Graham Harman in his (2010) and elsewhere.

Of course we do not spend our lives fully absorbed in rote activities in which our tools function smoothly. Sometimes the tools break down, or go missing, or instead present themselves as the wrong sort of thing (the form in front of me asks me to use a pen, but I have only a pencil). In these cases I am snapped out of my absorption with the task at hand and am forced to repair the tool, find it, or replace it. When my absorption is interrupted, the things before me stand out *as* objects, as *Vorhanden*, or present-at-hand. Instead of interacting with the pencil circumspectly, I now see it as a what it is—a dull object that needs to be sharpened—and, having sharpened it, I quickly return to work and the pencil withdraws again. The breakdown may be more serious—something I normally take for granted might vanish, for example, or suddenly stop functioning. And then I become aware of the entire context, the referential whole *within* which I have taken the thing for granted. Of course this does not mean that I can only think about entities explicitly when they break—I can do so at any time, as Heidegger makes clear.[48] The point is only that things appear to me *as* things, *with* purposes, and *of* a certain material nature only when I am not fully absorbed in using them.

Contemplative attitudes toward things—the sorts of attitudes which see things explicitly in physical or metaphysical terms—involve a stepping back from the normal context of absorption, in which our taking care of things is "subordinate to the in-order-to constitutive for the actual useful thing in our association with it." (69) Explicit thought about something, in other words, requires us to leave the agential stance in which we simply *use* the thing for its standard purpose. But obviously this does not mean that contemplation breaks free of the normal contexts of interaction with innerworldly beings;

[48] In *History of the Concept of Time*, he articulates three, rather than two ways of seeing entities, and he makes it clear that each of these can be entered into at any point.

it means only that such contexts are implicit in every act of explicit contemplation. To contemplate a pencil—to *know* that it is a pencil, I must already have access to the referential context in which that pencil has relevance. And since every reference in this context will point to something further outside of it, I can only grasp the pencil on the basis of a familiarity with the totality of references, i.e., the world. Thus, "being-in-the-world signifies the unthematic, circumspect absorption in the references constitutive for the handiness of the totality of useful things. Taking care of things always already occurs on the basis of a familiarity with the world." (76)

Of course if a grasp of the world is presupposed in any explicit taking care of things, this means that we can never make the *entire* context constitutive of a thing explicit. Dreyfus has made good use of this thought over the years, especially in his argument that it is impossible to get computers to *think*, or to interact with entities in the way we do, by teaching them a finite set of explicit rules. No set of explicit rules can fully articulate the underlying referential context. But in the debate with McDowell, Dreyfus takes the claim further: he argues that our taking care of things is not, at bottom, conceptual, but that it involves an altogether different *kind* of interaction and a fundamentally different sort of *content*. But both of these claims go far beyond the former claim that the world cannot be made conceptually explicit. It is one thing, in other words, to say that we cannot make the entire referential totality explicit all at once, and to say that the totality contains elements that cannot be made conceptually explicit. And, we might recall, McDowell's claim is simply that our perceptual openness to the world involves pre-conceptual elements, that is, elements that *can* be conceptualized even though initially one may lack—and may in fact always lack—any concept for them. I

think McDowell's account, then, comes far closer to Heidegger's than Dreyfus would have it. I will return to this point.

First, I want to address two apparent puzzles that Heidegger's account seems to raise. On the one hand, one might wonder whether his tool analysis is really an account of human agency, or of only a segment of it—that is, the segment that involves writing, hammering, and other uses of tools. On the other hand, Heidegger seems to objectionably assimilate all innerworldly entities to tools. What, one might ask, are we to make of trees or of the sun? Do those entities only appear to us when we need lumber or a way to tell the time, as Heidegger sometimes seems to imply? The answers to these puzzles naturally belong together. All human action is action in the world.[49] The point of Heidegger's analysis—or at least one of the points—is just that we often use tools without recognizing them as such. Tools are not just things like hammers and pencils. They are also sidewalks, that we use to get somewhere; or the sun, which we use for warmth, for light, or to get a tan. We may make use of trees not only for lumber, but also for atmosphere, or for entertainment; or we might treat them (like Hansel and Gretel) as something frightening to escape from. But even here the trees are, in some sense, "useful things," in that they make themselves manifest within a referential whole. Thus, the world in which we act is a world of tools, and human action almost necessarily involves making use of tools (in this sense) in some way. This is not a complete account of agency—we still have not gotten to Heidegger's explanation of purposes—but it is a largely comprehensive one.

[49] We can list "mental actions"—acts of thinking, deciding, realizing, and so on—in a special category, since these do not necessarily make use of things in the usual sense. But even that is not clear. One thinks or has realizations about *something*, something that is perhaps not an innerworldly being, but stands in some relation to such a being. And acts of deciding do, after all, involve deciding to do *something*, something that typically involves bodily movements that change something in the physical configuration of the entities around me.

The foregoing account of our being-in-the-world as a background condition of agency as well as any explicit thought—including CTD—must still be supplemented. First, the in-order-to constitutive of tools, and to which our activity in the world is subordinate, must have a source: our ways of using tools and of acting are appropriated from "the they," a topic I will address in the next section. Second, it is clear that the notion of purpose—that which makes the referential framework of world possible—has not yet been explained. Heidegger addresses this topic—which he calls an analysis of being-in as such, or the being of the there (133)—through three structural features essential to Da-sein (existentiales): attunement, understanding, and discourse. These features are equiprimordial; none of them have priority over the others, and all are intertwined, so that understanding is always attuned, and so on.

Attunement refers to our moods. While some Heidegger commentators maintain that the moods involved in attunement should be understood as distinct from emotions and perhaps even as "deep" moods that differ from surface moods, this does not seem to be what Heidegger intends. Instead, he explicitly refers to Aristotle's work on the passions, and the later philosophical work on "the affects and feelings," which "fall thematically under the psychic phenomena, functioning as a third class of these, mostly along with representational thinking and willing. They sink to the level of accompanying phenomena." (139) In condemning this tradition of relative neglect, Heidegger clearly does not single out moods as a special category distinct from what philosophers have addressed as affects and feelings; rather, he chides the tradition for losing sight of the importance of this "third class" of "psychic phenomena." That Heidegger uses a common emotional state—fear—as a key example of a mood reinforces this point.

But moods are not, on his view, subjective states that color a pre-given objective world. Rather, they are constitutive of our having a world at all—they disclose Da-sein in its thrownness. In other words, moods disclose the world in such a way that it can matter to Da-sein (or, better, in such a way that it already matters to Da-sein); it is thanks to them that regions of the world can have salience or, in Dreyfus's terminology, that affordances can solicit us. "Being-in as such is existentially determined beforehand in such a way that what it encounters in the world can *matter* to it in this way. This mattering to it is grounded in attunement, and as attunement it has disclosed the world, for example, as something by which it can be threatened." (137) On the other hand, moods disclose how Da-sein is in this world—how we relate to the world in which we find ourselves. Whether the mood is love or fear, it always simultaneously discloses two poles: it discloses the world in such a way that a region of it is loved or fearful; and it discloses Da-sein as that which relates to this region through love or fear. To have a world is to have regions of it stand out in this way—as something that attracts or repulses us, for example, that draws us in and makes us pay attention to it. My self-apprehension is thus linked to my apprehension of a world in which I find myself through moods.[50]

Understanding, on the other hand, adds the key component that Dreyfus mentioned earlier—identity. Like attunement, understanding involves co-constitutively revealing Da-sein and its world, and it reveals both by projecting possibilities. Just as affectivity discloses entities in the world as salient, understanding projects possibilities that are co-constitutive of both Da-sein and the entities with which it interacts. "In understanding a context of relations, Da-sein has been referred to an in-order-to in terms

[50] Although all moods reveal both Da-sein and its world, this does not mean that all moods reveal certain *regions* of the world as mattering. *Angst*, as we will see later, is a mood that reveals the world, but in such a way that nothing in it matters.

of an explicitly or inexplicitly grasped potentiality-of-its-being for the sake of which it is." (86) In using entities, in other words, we must already encounter them as having purposes—the in-order-to, or the task in terms of which such entities are useful. But such purposes are only meaningful for the sake of something else. Less abstractly, the point is that the tasks for which we use tools are, ultimately, defined by human purposes. In a stock example, one encounters a hammer in hammering boards together, and one hammers boards together in order to build a house, but one builds a house for the sake of an entity that dwells in houses. Similarly, one encounters the earth beneath one's feet as a "tool" for walking on; and as such it is meaningful within a framework ultimately circumscribed the common human need for transportation. The for-the-sake-of-which, then, gives the purposes in terms of which we grasp the referential whole within which entities are disclosed. And this means that Da-sein itself is disclosed—usually only implicitly—in everyday dealings with objects.

But the for-the-sake-of-which is not simply that in terms of which we understand entities: it is also that in terms of which we understand ourselves. Da-sein never, on Heidegger's view, understands itself exclusively in terms of its occurrent features, such as its height, age, nationality, and so forth. Rather, these features themselves are grasped only relative to Da-sein's potentiality-for-being, that is, it's projection of a possible way to be. And possibility, in this sense, is not just a logical or metaphysical possibility (in the sense that it is possible that one might—or might not—lose one's hair, or even sprout wings and fly off into the sunset), but rather Da-sein's competence in dealing with its world. "We sometimes use the expression 'to understand something' to mean 'being able to handle a thing,' 'being up to it,' 'being able to do something.' In understanding as an

existential, the thing we are able to do is not a what, but being as existing… Da-sein is not something objectively present which then has as an addition the ability to do something, but is rather primarily being-possible." (143) My having or lacking hair (or wings), my skill in playing chess, or my prowess with a samurai sword are objective facts about me, much the way that hardness is an objective fact about tables. But I *understand* myself in terms of these things not because I take them as defining features of myself, but because I take them up in the ways of being that I project for myself—that is, in my possibilities. Of course I *can* define myself by my kendo skill, but this will be because I already understand myself as a braggart about my sword abilities or, perhaps (in a different time and place) because I am honored to be able to slay the enemies of my shogun.

Thus, my self-understanding involves a projection of a possibility that I strive to fulfill or press into, and it is in light of these possibilities, our for-the-sake-of-which, that our occurrent properties (and possibilities, in the standard, non-Heideggerian sense) matter. To put it another way, "the project character of understanding constitutes being-in-the-world with regard to the disclosedness of its there as the there of a potentiality of being." (145) And it is always in terms of our possibilities that we understand ourselves—which is to say, as well, that our being is characterized by possibilities. Crucially, Heidegger holds that understanding always projects possibilities *as* possibilities: the possibilities in terms of which we understand ourselves are never attainable.[51] To attain or fulfill a possibility would be to make it into an occurrent characteristic—for example, I might project the possibility of becoming a kendo expert

[51] Blattner (1996), in an excellent analysis, refers to this as "the Unattainability Thesis: Dasein's ability-characteristics [in terms of which it understands itself] are not attainable." (107)

and then set out to master the art of kendo. But if I succeeded in mastering kendo, it would no longer be a *possibility*. To say that Da-sein projects (and, therefore, *is*) its possibilities *as* possibilities is to say that the possibilities in terms of which Da-sein understands itself are not things like mastering kendo, but *being* a kendo expert, and one can certainly have expertise in kendo without understanding oneself as a kendo expert, since one can simply—despite having mastered kendo—see oneself as a lawyer who happens to also be a kendo hobbyist. This is why Da-sein's for–the-sake-of-which is tied to its *potentiality* of being: Da-sein always exists *as* possibility, or *as* potential, and not as actuality or fact.

The possibilities we project for ourselves are thus co-constitutive with the possibilities in terms of which entities in the world are disclosed to us (my self-understanding of myself as a kendo expert is intimately tied to my understanding of the kendo sword). "The essential possibility of Da-sein concerns the ways of taking care of the 'world' which we characterized, or concern for others and, always already present in all of this, the potentiality of being itself, for its own sake." (143) Commentators who focus on the connection between our interactions with objects and our identity sometimes focus on overly particular identities (e.g., the way I handle a hammer might reflect my identity as a carpenter) and argue that using entities appropriately—following certain norms—involves simultaneously intending oneself as a certain kind of person (that is, in accepting the correct use of a hammer, I *co-intend*—and thus constitute—myself as a carpenter).[52] But the way Heidegger's account seems to work is both simpler and more complex than this. Consider the following set of considerations: As an embodied being, I am sometimes tired. It is customary in our culture to sit when tired. It is also customary to

[52] This is the way so-called pragmatists tend to read Heidegger. See, e.g., Okrent (2000a, 2000b).

sit in class, during job interviews, and so on. And my world includes various objects, some of which are designed for sitting on. Others are merely convenient, and yet sitting on them is not ruled out by the social conventions applicable within particular situations (though it may be ruled out in others). And this set of considerations—along with many others, which it may be impossible to fully articulate—is what allows me to interact with (or take care of) certain regions of my world as chairs or, more generally, as seats. Furthermore, the way I sit, my posture, the movements involved in lowering my body, indicate my acceptance or rejection of (and, in either case, my background awareness of) social conventions, my sense of comfort or discomfort, the shape and hardness of the object I am sitting on, and so on. And, we might add, I can take care of the chair for other purposes. Perhaps a light bulb has gone out. The lack of a ladder, the height of the chair, my need to change the light bulb, the fact that I am taller than my household partner or perhaps simply the sense that I should be doing more around the house—in light of these considerations, I might see the chair as a surface to step on to get closer to the ceiling.

The point of this long-winded discussion of seats, chairs, and step-stools is just this: I can interact with something as a chair only in light of a totality of considerations through which the chair becomes significant. But unpacking these considerations is difficult. Some of them refer to the materiality of the object. Others refer to social norms, or to my standing with regard to those social norms. Yet others—including considerations of my standing with regard to social norms—refer to my identity: for example, I am someone who needs to do more around the house, I am someone who occasionally becomes tired (and is not embarrassed to let others see this), and so on. Part of the story here is that the various sorts of considerations—relating, respectively, to

entities unlike Da-sein, to our social dimension (the they), and to Da-sein itself—are intertwined in such a way that picking them apart requires standing away from them and seeing them as referring only in a specific direction. And part of the story is that what is meant by Da-sein's "identity" can be a far more everyday and banal matter than a profession, which is the standard example used by commentators. In fact, it must be, since my identity as, say, a philosopher could not possibly account for the myriad ways in which entities in the surrounding world appear to me in my taking care of them. Identity is dispersed among different self-conceptions, different social norms, and the innerworldly entities to which it is related.

Dreyfus is right to reject the pragmatist reading by pointing out that, on Heidegger's view, "a role or identity organizes all of one's activity. One does not have an identity because one acknowledges tool using norms as Okrent claims, but one uses tools and people, normally or idiosyncratically, in order to manifest one's identity… Self-reference is not a feature of each act; it is the way many of one's actions are organized or coordinated." (Dreyfus 2000b 341) Da-sein does not constitute itself—explicitly or implicitly—every time it follows a norm in its taking care of things. Instead, Da-sein's identity—that for-the-sake-of-which it acts—provides the clearing within which innerworldly entities manifest themselves to be used (together with their norms) in the first place. But we should avoid falling into the trap of thinking that Da-sein must explicitly pick an identity before it can have a world. The opposite seems to be the case: "As essentially attuned, Da-sein has always already got itself into definite possibilities… But this means that Da-sein is a being-possible entrusted to itself, *thrown possibility* throughout… And since understanding is attuned and attunement is existentially

surrendered to thrownness, Da-sein has always already gone astray and failed to recognize itself. In its potentiality of being, it is thus delivered over to the possibility of first finding itself again in its possibilities." (144) It is true, in other words, that Heidegger sees our for-the-sake-of-which as structuring our interactions with entities in the world. But precisely because the two go hand in hand, Da-sein gets lost in those interactions. Rather than using tools "in order to manifest one's identity," Da-sein has to discover its identity as something with which it is already saddled, and which it already manifests; that is, Da-sein can find itself in its taking care of things because that taking care is structured by its projected self-understanding, but it finds itself *first* (or, in Heidegger's terminology, proximally and for the most part) among the things it takes care of.

The intertwining of attunement and understanding, of thrownness and projection, is crucial to Heidegger's account: I find myself *in* possibilities; I project on the basis of the circumstances in which I find myself. Blattner illustrates this relation with the example of a decision to become a lawyer as the projection of a possibility. I might, for example, find myself associating law with power, and it is because I already have this association, and because I already care about power, that law appeals to me:

> Affectivity lets possibilities show up in determinate ways, as mattering to me in determinate fashions. I already care about power, or money, or helping others, or whatever, and this already caring guides my decisions… The possibilities themselves show up for me in the light of the affectivity: law seems worthwhile to me, for I care about power…. My attunements are the grounds for my action; they make my projections possible. I act on the basis of my attunements… Thus, attunement is essential to projecting oneself into possibilities. (Blattner 1992 108)

This is right, but since attunement and understanding are *co-constitutive*, the relation works the other way as well: my projecting self-understanding modifies the meaning of

the situation in which I find myself. As mentioned earlier, occurrent features of Da-sein are never simply givens; they are not factual, but *factical*, in that their mattering to me is always taken up within my for-the-sake-of-which, and takes its meaning from the possibilities I project. In the case of the entities I take care of, their materiality both places limits on, and is disclosed within, their uses. Similarly, Da-sein's thrownness both saddles it with possibilities and itself becomes meaningful in light of those possibilities. As is well known, Heidegger identifies Da-sein with care, with the care structure incorporating both our thrownness and projection, our encounter with entities in light of our projection of world and identity: "The being of Da-sein means being-ahead-of-oneself-already-in (the world) as being-together-with (innerworldly beings encountered)." (192)

Finally, this structure strongly suggests that one cannot separate our taking care of entities from our identity in the way Dreyfus suggests. Our identity—the for-the-sake-of-which—is co-extensive with the world within which we take care of things. Our taking care, then, cannot operate as a background condition detached from our identity. Nor can one say in response to McDowell—as Dreyfus does—that our coping activity is not rational because "most of our activities don't involve concepts at all. That is, they don't have a situation-specific 'as structure'". (Dreyfus 2007b 371) What Dreyfus says here is partially right—depending on which "as structure" he has in mind. To clarify: Heidegger presents interpretation as a development of understanding. But he is careful to note that interpretation, in this sense, is not yet linguistic. In interpretation, entities are made explicit by being disclosed *as* something: "The circumspect interpretive association with what is at hand in the surrounding world which 'sees' this *as* a table, a door, a car, a

bridge does not necessarily already have to analyze what is circumspectly interpreted in a particular *statement*." (149) Instead, statements that present something *as* something are founded on this more primordial interpretation, and they are deficient in their disclosive capacities. As we've already noted, an entity is disclosed against the background of a referential whole; language, as such, necessarily detaches the entity that it discloses from that whole. Heidegger therefore distinguishes the hermeneutic *as* of interpretation from the apophantic *as* of the statement.

The distinction between two types of *as* structures indicates that there is a pre-linguistic *as*. Thus, Heidegger points out that a hammer can be disclosed as too heavy within interpretation but without stating that it is too heavy. Circumspective interpretation "may take some such form as 'the hammer is too heavy' or, even better, 'too heavy, the other hammer!' The primordial act of interpretation lies not in a theoretical sentence, but in circumspectly and heedfully putting away or changing the inappropriate tool 'without wasting words.'" (157) Thus, the *as* structure of interpretation, instead of stating that some property belongs to some entity, treats the entity as having that property precisely by dealing with it in a certain way. Dreyfus, oddly, wants to detach interpretation from our absorbed coping experience, insisting that interpretation only enters onto the scene "when we are no longer able simply to cope." For example, "when the doorknob sticks, circumspection discovers what the doorknob is for, although it fully understands it only in using it." (Dreyfus 1991 196) But this seems to be the exact opposite of Heidegger's claim that "any perception of useful things at hand always understands and interprets them, letting them be circumspectly encountered as something." (149) If interpretation were involved only when we cannot continue to cope, it could hardly be present in all

circumspective dealings with entities, as Heidegger suggests. The point, instead, seems to be that we already interpret the doorknob—even when it is functioning perfectly—by treating it as a doorknob, that is, by turning it.

It is true, of course, that most of our dealings with things—on Heidegger's account—do not utilize concepts at all. But this fact is perfectly in accord with McDowell's positing of a pre-conceptual domain: that is, the idea, developed above, that whatever we encounter in the world already is conceptual, not because we have already conceptualized it or even have the words with which to do so, but because it is of the right *form* to be taken up into concepts. And this is precisely the view Heidegger suggests, when he notes that *fore-conception* is one of the major structures of interpretation. That is, our understanding of the world as a referential whole allows us to interpret—i.e., to deal with entities in the world as such and such—and interpretation, in turn, provides the structure taken up in explicit linguistic conceptualization. Our concepts are founded in a pre-conceptual understanding of world; but this does not mean that the world as such is a non-conceptual background: the world *as a whole* cannot be conceptualized; but it provides the form that allows for conceptualization. Thus, Heidegger's view gives us something similar to the McDowellian picture, with the main distinction being a matter of emphasis: McDowell stresses that conceptuality pervades all our activity, whereas Heidegger instead stresses that our activity is the basis on which conceptuality operates. There is, among critics of Heidegger and even among some supporters, a tendency to read him as an irrationalist, but it is interesting to see what he himself says about the traditional philosophical view of human beings as "*zoon logon echon*": "The later interpretation of this definition of human being in the sense of the

animal rationale, 'rational living being,' is not 'false,' but it covers over the phenomenal basis from which this definition of Da-sein is taken. The human being shows himself as a being who speaks. This does not mean that the possibility of vocal utterance belongs to him, but that this being is in the mode of discovering world and Da-sein itself." (165) In other words, Da-sein *is* properly understood as rational, but its rationality must be seen as implicating the entirety of the care structure.

We can confirm this point by remembering that Heidegger lists discourse as equiprimordial with attunement and understanding. Discourse is, on his view, "the existential-ontological foundation of language" (160-161)—it is the existential characteristic of Da-sein that allows it to communicate, and to communicate *about* something. Thus, discourse articulates the world in accordance with an understanding (allowing for a grasping of that understanding in interpretation) and an attunement (which in language is brought out through the rhythm or a way of speaking). Discourse allows us to have a shared world, which we can express to each other and make ourselves understood. And Heidegger provides evidence that discourse underlies language in the fact that we can understand—to some extent—what is expressed by someone speaking a language we do not know. As equiprimordial with attunement and understanding and, therefore, co-constitutive of world disclosure, discourse must also be operative in all circumspect taking care of the world. The entities we deal with must be entities we can make someone else see, communicate something about. And again, that discourse is not conceptual does not support Dreyfus's view, but rather McDowell's: what we can communicate to others must have a form by means of which it *can* be taken up

conceptually, though it need not be, and though sometimes in practice one might find it impossible to do so.

To conclude, let me now make a rough terminological distinction, which I will later rely on. I have been arguing that the background on which we act is such that it can be taken up and articulated rationally, so that even if one does not explicitly deliberate about what one is doing, and even if one is utterly absorbed in a task, one may still be acting for a reason. Let me now call the background *character*. One's character is what, for the most part, one acts on; for example, when one responds to solicitations, they are soliciting our character. Character as articulated through discourse, on the other hand—that is, not as full-blown linguistic self-understanding, of the sort used in CTD—I want to call *will*. So using this terminology we can say that agents have both a character and a will. If I speak of someone driven to act by his love, I am referring to his character. If I speak of him as acting because his action serves the interest of his beloved (for example), I am referring to his will. Or, to take another example, to say that someone is honest might mean that he *cannot* lie, is not suited to it, and then one is referring to his character. Or it might mean that he appreciates truthfulness, and acts on that reason. And this is his will. So we might think of Dreyfus as attempting to separate character from will, whereas McDowell's—and, I believe, Heidegger's—view reconciles them. They refer to the same content, that is, the background involved in the agent's agency. But they refer to it differently conceived: as a "natural" trait, or a rational one (second-nature, as McDowell calls it). As exhibiting a pattern of actions, or as exhibiting a pattern of thought. This terminology allows us to say that on the attributionist view, when the agent acts

according to his character, he is also expressing his will. It is this latter element that makes his action of a type that can, in principle, be said to be in his control.

B. History and Ownership

My aim in the above was to sketch out what is involved in an account of acting rationally without CTD. Because attributionists take such cases of absorbed coping as paradigmatic in their account of responsibility, it is important to see just how the idea that evaluative judgments or values are embodied in our pre- or non-deliberative actions and attitudes is supposed to work. And, as I argued in the last chapter, the important point is not just that much of the activity for which we typically hold people responsible is non-deliberative, but that any volitionist account of responsibility must still be grounded in pre-deliberative processes (e.g., the processes on the basis of which we take up some considerations rather than others in the deliberation itself, and the processes on the basis of which some considerations even *count* as considerations within the deliberation). Making sense of pre-deliberative agency is thus crucial to working out how anyone can be responsible in the first place.

In following a construal of Sher's argument that no semantic component need be involved in the connection between features of the agent's character and the agent's actions, I tried to work out, by means of the McDowell-Dreyfus debate, a way of making sense of the thought that our actions may proceed from a pre-deliberative background and yet still be rational, and thus subject to agential control. The sense of rationality involved turned out to be not one where our agency is maxims all the way down (in Dreyfus's

critical formulation)—so that every human action or attitude is the upshot of a prior rationalization—but one where rationalization and conceptualization is always *retrievable*. Agents can act for reasons even in the absence of CTD because they can articulate the reasons for their actions retrospectively and, even when they cannot, this need not serve as evidence that their action was arational. The processes involved in generating action may be rational, in other words, so long as the capacities operating in them are ones accessible to reason.

I proceeded to work out the basics of a Heideggerian theory of agency in order to suggest (indirectly) that we can thereby work out the most comprehensive account of how this notion of control might work, and especially of the crucial idea that rational considerations such as judgments or values are *embodied* in our actions and attitudes. This idea is crucial because it allows agents to be responsible for their actions and attitudes *directly*, without the mediation of prior thought, and this is just what we need to explain responsibility for pre-deliberative action. On a Heideggerian view, we might say that attitudes and actions reflect values or judgments of the agent *because* those attitudes and actions stem from the agent's care—that is, from her affectivity as it is modified by and taken up in her projected self-understanding, which constitutes her identity as an agent. In offering the Heideggerian account, to be sure, we transform the intent of attributionism, which is to claim that our rational judgments are (at least) partly constitutive of our attitudes and actions. Instead, our picture now presents those judgments—explicitly stated—as derivative of and thus deficient relative to the referential totality within which our attitudes and actions have their full significance. But this preserves the basic picture: the rational judgments are no longer constitutive of our

attitudes and actions, but they still reflect our selves insofar as they are derived from activity that reflects our selves since it is reflective of our for-the-sake-of-which.

At the same time, however, the Heideggerian account should make it easier to see that there is something deficient in the attributionist picture worked out so far. For one, we can return to the difference, already noted by Levy, between expressing and reflecting. On the Heideggerian picture, at this point, the agent's attitudes only *reflect* her self—they do not necessarily express it, since the agent might simply find herself thrown into her possibilities and have to *find* herself within them. The same seems to go for the attributionist account since, if anything, it is significantly less complex. In fact, on the attributionist account the connection between agency and the agent's self or will is far more tendentious. On the one hand, attributionism—as we have seen—is intended to correct a difficulty with Frankfurt's view, on which agents are responsible only for acting on those desires with which they identify. The attributionist, recognizing that an agent's self is wider than the narrow scope of identification, attempts something like a coherence theory of the self. But on the other hand, giving up the link with identification makes it difficult to see just how the various judgments involved are genuinely the agent's own, rather than simply belonging to her. Attributionists may attempt to fix this, as we have seen, by postulating criteria of integration and depth: an attitude is the agent's own, on this view, if the judgment it embodies is sufficiently well integrated with the agent's other judgments, and if it is especially difficult to change. But *this* may seem to bring us to the unappealing Stoic view that we can change our attitudes simply by changing our judgments; discover what is good or evil, and the attitudes will follow!

Consider two points here. First, an example: an individual may discover that he has strong sexual urges. These sexual urges are reasonably well integrated with his other attitudes (at least, considering how un-integrated most of our desires are): they are especially difficult to shake (and thus deep) and they are not opposed by other beliefs and values the agent holds. But it is possible that he simply cannot see a good reason to satisfy those urges; "it feels good" is not a justification he accepts. He does not see anything wrong (or at least seriously wrong) with those urges, so there is no significant opposition with other beliefs. But it is hard to see how there is any *valuing* on his part going on. Should we say that he does not, in fact, value sex, we are making his values depend on his endorsement, which is what the attributionist account is meant to avoid. On the other hand, if we insist that—since the relevant attributionist criteria are fulfilled—the man genuinely *values* sex, the claim seems awkward. We may just as well claim that his *body* values sex and makes evaluative judgments with which *he* disagrees (or towards which he is simply indifferent).

Second, the difficulty with the attributionist criteria is that the criteria are entirely third-personal. This is not to say, of course, that it is especially easy to determine how integrated an attitude is from the outside; the point is that the criteria are specifically designed to be open to external evaluation. Perhaps the agent himself is in a privileged position with regard to knowing his desires and beliefs, but he is in a better position to know whether or not he is responsible than an outside observer only by virtue of this epistemic fact. The desires and beliefs relevant to judgments of responsibility seem here to be only properties that an agent *has*; since he need neither identify with nor choose them in any sense, they belong to him in the same way as any external property—like his

height, or his shoes.[53] That this is the attributionist view is especially clear on Arpaly's account: she likens moral criticism to the sort of criticism involved in saying that someone isn't good at business. Attribution of responsibility is viewed as just another kind of third-personal description. Arpaly notes that, in the case of business, one can make a judgment that someone is good or bad at it, and the person's history—what made her that way—is completely irrelevant to the judgment. Moral criticism, on her view, is descriptive in the same way: whether or not your bad will is traceable back to a bad childhood or something of the sort is simply irrelevant to the quality of your will now, and praise and blame are judgments that respond to one's quality of will.[54] But this view strikes me not as misrepresenting the phenomenon of responsibility, but as missing it entirely. To say that what makes blame warranted or unwarranted is the quality of an agent's will is to *presuppose* that the agent is responsible for her will in the first place.

The idea is this: responsibility involves being praiseworthy or blameworthy (that is, being an appropriate target for praise or blame). For the attributionist, then, if an agent satisfies the criteria for deserving praise or blame (on their view: having a good or ill will), he is thereby shown to be responsible. My point here is that the reverse holds: if an agent can *fail* to be responsible and yet manifest a good or ill will, praise and blame will

[53] This, of course, is cognate to Levy's criticism.

[54] This comparison of moral criticism with criticism of business ability only seems wrong, she thinks, if one holds praise and blame to require or be akin to punishment and reward, as something that can be required or forbidden, or as something appropriate or inappropriate. Instead, she argues, "it is first and foremost *warranted* or *unwarranted*, the way that my fear of getting a flu shot is warranted only if flu shots are dangerous to me." (Arpaly 2003 172) I doubt, however, that *this* is the real problem, or that we can make much sense of the idea implied here, that praise and blame are the upshots of epistemic judgments. In any case, not all attributionists share this view—the important point, as we've seen in Angela Smith, is that the attributionists hold that anti-volitionism should be easier to swallow once we uncouple praise and blame from reward and (especially) punishment.

be inappropriate. Recall that the attributionist's aim is to give criteria for responsibility that justify our actual practices of praise and blame:

> I have treated moral blame as justified when one person correctly judges that another has guided his actions in a way that expresses contempt or disregard for the first person's moral standing. The justification of blame, then, has mainly to do with our capacity to guide our actions in a way that reveals our attitudes toward others, and this requires no investigation into how a wrongdoer came to possess the dispositions that incline him to exercise his power of self-governance as he does. (Talbert 2009 18)

So what is supposed to justify judgments of praise and blame is that the agent's actions reveal her attitudes—that is, whether her attitudes guide her actions (or, in Smith's case, whether her values guide her attitudes; there are clearly a few variations on this theme). If they do—if the proper reasons-responsive mechanisms are in place—then we can simply hold the actions (or attitudes) blameworthy because they reflect underlying blameworthy attitudes (or values).

The fact that we praise or blame someone implies that we hold the agent responsible. But it does not follow that the agent *is* responsible. A crucial premise needs to be inserted, tying the *agent* to the quality of her will. As the last paragraph demonstrates, attributionists are already divided over what is supposed to be representative of the agent: her attitudes or her evaluative judgments? That attributionists disagree with each other, of course, does not show they are all wrong. What it *does* suggest in this case, however, is that they take praise and blame to be justified by factors that are stand-ins for the agent: we blame agents because their attitudes or evaluative judgments express who they are. And these features are supposed to express who the agents are because they—as expressions of rationality—are representative of the agent as a whole. The problem is that this does not follow. Agents are not automatically

identifiable with each of their rational operations, nor necessarily even with the coherent bulk of their rational operations; they are identifiable only with those operation that are *their own*. And ownership is a trickier feature than attributionists allow.[55]

Aside from Frankfurt's developing view of identification and the attributionist view of integrity or coherence (provided we can take these as attempts to explain ownership), one of the most significant recent accounts of ownership is provided in Fischer and Ravizza's *Responsibility and Control*. In earlier work, Fischer had already developed what he calls semi-compatibilism, that is, the view that moral responsibility is compatible with determinism (although free will—in at least one sense—is not). There, he had argued that regulative control—which involves alternative possibilities—is incompatible with determinism. Following Frankfurt, however, he argued that moral responsibility does not require alternative possibilities and can therefore be made compatible with determinism. What is needed for moral responsibility is guidance control, which requires that agents act on reason-responsive mechanisms. Response to that early work often argued that there was a problem with Fischer's account of reasons-responsive mechanisms: such mechanisms could conceivably be implanted in agents through manipulation.[56] Aside from strengthening the account of reasons-

[55] In discussing whether we should hold agents who have been the subjects of manipulation responsible, Talbert seems to address this concern: "One way to put the general point here is to say that the question we should ask when confronted with a manipulation scenario is not really whether the values that an agent now has are *her* values – in a sense, values cannot fail to be those of the agent who acts on them. Rather, …we should ask whether her actions issue from the right sort of internal states such that they are capable of expressing interpersonally significant values, attitudes and judgments about reasons." (Talbert 2009 13) But to say that "in a sense, values cannot fail to be those of the agent who acts on them" is simply question-begging. As the phrasing implies, and as I will argue later, there are at least two different senses in which values can belong to an agent; which sense we have in mind will be relevant to whether "her actions issue from the right sort of internal states."

[56] Manipulation scenarios often involve science fiction evil scientists poking around in someone's brain, often disregarding mental holism and anti-reductionism at will. The point, for the most part, is both to clarify the conceptual field of moral responsibility and to provide scenarios similar enough to determinism

responsiveness—a topic I will not address here—*Responsibility and Control* lays out a new, historical condition meant to allay manipulation concerns: guidance control now requires both that the mechanism that issues in an action be reasons-responsive, and that it be the agent's *own* mechanism.

In introducing their account of ownership, Fischer and Ravizza stress that they have in mind a *historical* notion of responsibility: agents must *take* responsibility for their reasons-responsive mechanisms at some time, and they thereby make themselves responsible for any actions that issue from these mechanisms in the future. What makes the account historical, then, is that responsibility for an action requires a past act (broadly construed) of taking responsibility on the part of the agent. To set up the account, Fischer and Ravizza contrast historical phenomena, which depend somehow on their history, with nonhistorical, "current time-slice" phenomena, which depend only on "snapshot" properties. (Fischer and Ravizza 1998 171) For example, the property of being a correct answer to a math problem does not depend on the history of how someone came to that answer. And a book's property of having exactly 173 pages does not depend on how those pages came to be in the book; it depends only on the number of pages the book has at a particular time. On the other hand, some phenomena are historical. A simple example the authors give is that of being a genuine Picasso. No snapshot properties of a work—that is, no properties present here and now—can determine whether a painting is a genuine Picasso; that depends entirely on whether or not it was painted by Picasso. (Of course snapshot properties—carbon dating, brushstroke, type of paint, etc.—can be used by art historians to ascertain whether a painting is indeed a genuine Picasso, but those

---

as to raise doubts (if one allows that the manipulation in question rules out moral responsibility) about compatibilism. For a discussion of manipulation-based arguments against Fischer's early account, see Ekstrom (2000 169-173).

properties clearly do not *make* the painting such.) Another example is drawn from Robert Nozick's view of justice. Roughly, the idea is that we cannot determine whether a given distribution of resources is a just one without knowing the history of that distribution. For example, it is possible that most people living with a just distribution chose—according to just transactions—to give their money to a particular individual in that society, creating massive wealth inequality. The existence of inequality, however, does not show that the current distribution is unjust, since there was no injustice involved in its coming about. On the other hand, exactly the same distribution could be brought about through theft, which would make it unjust.

The argument that responsibility is a historical notion, then, claims that there are certain conditions prior to an action that must be met in order for the agent to be responsible for it. To demonstrate why a historical view of responsibility is needed, Fischer and Ravizza contrast their account with what they call "mesh" theories of moral responsibility—theories that require a mesh between some features of the agent. (Fischer and Ravizza 1998 183-186) They give three examples. For Frankfurt, as we have seen, responsibility requires a mesh between the agent's first-order desires and his second order volitions—this constitutes identification with the first-order volition on which the agent acts and is supposed to be sufficient for moral responsibility. Watson's account, similarly, requires a mesh between an agent's desires and her values. And, finally, attributionist theories are clearly mesh theories of this sort as well: they require a mesh between actions or attitudes, on the one hand, and the agent's character, on the other.

But the problem with such theories is that they are indifferent to *how* the mesh came about: what matters for responsibility is whether or not the mesh exists. But,

Fischer and Ravizza contend, there are ways of creating such a mesh that are intuitively responsibility undermining. Aside from the fictional manipulation scenarios, hypnosis and brainwashing may also bring a mesh into being. Anecdotal evidence suggests, for example, that it is possible to quit smoking through hypnosis—those who undergo it find that they identify with their desire not to smoke rather than their first-order desire to have a cigarette. And certainly prominent cases of brainwashing are easy to find. Thus, mesh theories seem to falter because they do not require that the key elements in the agent be his own—identifications acquired through hypnosis are, intuitively, not one's own, and this seems to undermine the agent's responsibility.[57] One possible conclusion to draw from this, as I suggested earlier, is that mesh theories alone are insufficient—to be responsible, an agent must have control over whether or not the mesh exists. But Fischer and Ravizza take a different approach: they argue that the problem with mesh theories is that they require a mesh between current time-slice properties of the agent, when what is in fact needed is a mesh between temporally distinct elements.

To see what is needed, Fischer and Ravizza describe the process by which one becomes a moral agent, which involves three interrelated components: training, taking responsibility, and being held responsible. That is, children are typically trained to take responsibility for their reasons-responsive mechanisms, and are then held responsible (and hold themselves responsible) once the training is concluded. The training is needed in order to bring about the process of taking responsibility, and being held responsible is important in seeing that the training has worked—that is, someone who genuinely cannot understand the reactive attitudes others take toward him (regardless of whether or not he

[57] Of course agents can be responsible for the results of hypnosis, brainwashing, or manipulation if they undergo these voluntarily, as do the smokers who use hypnosis to help them quit.

agrees with them) has not yet become a moral agent. But the middle component—taking responsibility—is clearly the most important one, as "the process by which an agent takes responsibility for the springs of this action makes them *his own* in an important sense." (Fischer and Ravizza 1998 210) Taking responsibility, in turn, also consists of three components.

First, individuals must learn to recognize themselves as agents; that is, the individual "must see that his choices and actions are efficacious in the world. The agent thus sees that his motivational states are the causal source—in certain characteristic ways—of upshots in the world." (Fischer and Ravizza 1998 210-211) Next, they must learn to see themselves as apt targets of reactive attitudes on the basis of their exercise of their agency: not only do their motivational states have effects in the world, but those effects can be fairly subject to praise and blame. This need not, the authors stress, involve any metaphysical deliberation about fairness. Rather, "the individual must see that in certain contexts it is 'fair,' in the sense of being part of our given social practices, for others to subject him to the reactive attitudes in certain circumstances. That is, he must see that it is an appropriate move in the relevant 'social game' to apply to him the reactive attitudes in some contexts." (Fischer and Ravizza 1998 211) This step involves recognizing and understanding the reactive attitudes others take towards one's exercise of agency. That is, agents must learn how the game is played, and see themselves as players in that game.

Of course it is not enough to simply understand the game; seeing oneself as a player requires internalizing one's role. That is, agents must come to hold reactive attitudes towards themselves that correlate appropriately with the reactive attitudes others

take towards them. Fischer and Ravizza stress that the correlation need not be complete: it is perfectly reasonable for moral agents not to feel guilty in cases where others blame them, since one can clearly be a moral agent—in the sense of being an appropriate target of praise and blame—without fully conforming to social norms. An abolitionist who helps slaves escape may rightly refuse to feel guilty about the negative judgment his neighbors pass on him; the point is only that he must see the appropriateness of applying those attitudes to himself. So moral agreement with one's community is not necessary, but being a moral agent does require one to at least recognize the significance of the reactive attitudes that go along with violating the relevant norms. The authors thus picture taking responsibility as taking part in a conversation, where all sides can competently use the language of praise and blame even if they do not always agree on the cases in which the terms are to be applied.

The final ingredient in taking responsibility is that the individual must come to see himself as an agent and an apt target for reactive attitudes in an appropriate way on the basis of evidence. Normally, this means that agents—usually as children—figure out how their desires and actions impact the world, and are taught by their parents and other members of the community that some actions appropriately draw particular types of responses. The account is complicated, since it clearly involves acquiring normative beliefs on the basis of *evidence*, but once again, the agent need not have any deep appreciation for the normative grounds—he need only learn that certain norms apply, that they are deployed in particular ways, and that he is himself subject to them. The point of this condition is that the agent must take responsibility in ways that do not involve any responsibility-undermining kind of manipulation.

To be responsible for an action, then, agents must first undergo the entire process of taking responsibility outlined above. But to make the condition historical, Fischer and Ravizza insist that what agents take responsibility for are not their actions themselves at the moment they occur, but the reasons-responsive mechanisms that issue in those actions. Or, more precisely, "having taken responsibility for behavior that issues from a kind of mechanism, it is almost as if the agent has some sort of 'standing policy' with respect to that kind of mechanism. Thus, when the agent subsequently acts from a mechanism of that kind, that mechanism is *his own* insofar as he has already taken responsibility for acting form that kind of mechanism." (Fischer and Ravizza 1998 216) In particular, they mention two types of such mechanisms: our mechanism of ordinary practical reasoning, and our nonreflective mechanisms. Typically, children learn not only that they are apt targets of reactive attitudes when they deliberate about their actions, but that sometimes they can also be blamed (or praised) for actions that issue out of habit. This, of course, is supposed to explain how we can hold people responsible for the sorts of pre-deliberative (or non-deliberative) actions that have been the topic of my discussion above: since they recognize that those actions spring from their own agency and that they can fairly be blamed for them, they have taken responsibility for the mechanisms that issue in those actions and are thereby responsible for the actions.

This account helps to explain why manipulation rules out responsibility: in taking responsibility for their *ordinary* reflective and nonreflective mechanisms, agents do not also take responsibility for mechanisms that might later be implanted through manipulation, although it is possible after manipulation of some sort occurs for an agent to take responsibility for the actions that issue from this new mechanism. (This is why we

might hold the subjects of brainwashing responsible in some cases—Patty Hearst is one prominent example.) As this point clearly shows, Fischer and Ravizza's account of responsibility is, in their words, a "subjectivist" one. "In order to be morally responsible, a person must see himself as an agent who is an appropriate candidate for the reactive attitudes." (1998 229) Thus, a person can be responsible for an action only if she has taken responsibility for the mechanism that produced it; if she has not taken responsibility, on the other hand, she cannot be held responsible. To say that the account is subjectivist does not, of course, mean that it is *merely* subjective. That is, taking responsibility is a necessary, but not a sufficient condition for responsibility. Non-subjective conditions—such as the evidence-sensitivity involved in taking responsibility—are crucial as well.

This point, of course, raises an immediate objection: that people can opt out of being held responsible simply by refusing to take responsibility. But, as Fischer and Ravizza respond, while this is true, it is not clear that opting out in this way is possible voluntarily, or that it is at all desirable. First, to avoid responsibility for a particular action, an agent would have to not simply refuse to see herself as a fair target of reactive attitudes in the case of that action; she would have to fail to see herself that way with regard to the entire mechanism that produced the action, and this mechanism can stretch back into her childhood. So to avoid being responsible for any particular action that issues from one's ordinary mechanisms, the agent would somehow have to either go back in time and fail to take responsibility for the relevant mechanism in childhood, or she would have to have a complete breakdown sufficient for loss of responsibility prior to the action (this may be the strategy Hamlet was aiming at), but in such a way that she were

not responsible for the breakdown itself. Moreover, as Fischer and Ravizza contend, it is not obvious that agents can voluntarily refuse to take responsibility since, on their account, doing so involves not making a conscious decision, but something more like acquiring a cluster of beliefs, mostly in non-deliberative ways. And it is not easy to avoid acquiring beliefs, such as the belief that one can cause events in the world or that one is a player in a particular kind of game. And finally, even if one can voluntarily choose to avoid taking responsibility, the consequences would be dire. Failing to recognize oneself as the source of events in the world essentially robs one of any control over their actions; and failing to see oneself as an apt target of reflective attitude excludes one from the vast majority of interactions that make human life worthwhile—friendship and love, for example, are difficult to participate in for someone incapable of recognizing that another person's reactive attitudes toward them are at all meaningful.

While this account of taking responsibility has garnered a great deal of largely positive attention, it is not clear that it succeeds on either of the two fronts that I am concerned with here: demonstrating that responsibility is a genuinely *historical* phenomenon and providing an adequate account of ownership. Let's take up the first point. While history is introduced primarily to deal with various manipulation cases, critics immediately set out to work out numerous ever-more fantastic scenarios in which the process of taking responsibility is itself manipulated, but in such a way that the condition of appropriate evidence-sensitivity is not violated. (Haji 2000) If this is possible, then the historical condition does not succeed in providing an account that can withstand manipulation cases and—unless appropriate ways of fixing the problem can be

found—the motivation for giving a historical account of responsibility is seriously undermined.

Second, as we have seen, the process of taking responsibility is supposed to work as follows. An agent first recognizes himself as the apt target of reactive attitudes on the basis of particular behaviors. He *thereby* adopts something like a standing policy with regard to the *mechanism* underlying those behaviors, and this makes him responsible for any future behavior issuing from the same mechanism. But this is puzzling. If an agent adopts his standing policy on the basis of a finite set of behaviors, this is not intuitively sufficient to make him responsible for just *any* future behaviors produced by the same mechanism, especially since the agent need not (and cannot) know all the details of the mechanism when he takes responsibility for it.[58] But no one can predict every possible upshot of a mechanism, especially if one does not know the details of the mechanism.

To develop this point, we might keep in mind that a good deal of "situationist" psychology has strongly suggested that, often, environmental factors provide a far better explanation of actions than reference to agents' character (or reasons-responsive mechanisms) could. And even without the strong conclusions often derived from the research, it is fairly clear that our actions are at least strongly influenced by the context in which they take place—if they were not, after all, they could not be reasons-responsive at all. But if so, one might ask how an agent who has taken responsibility for a mechanism on the basis of actions in a limited range of contexts could take responsibility for what that mechanism might produce in an entirely new context. Many college students—to

[58] As Fischer and Ravizza concede, "in taking responsibility for acting from a kind of mechanism, one takes responsibility for acting from the mechanism in its full reality. To employ a metaphor, when one takes responsibility for acting from a kind of mechanism, it is as if one takes responsibility for the entire iceberg in virtue of seeing the tip of the iceberg." (1998 216-217)

take just one example—discover many new things about themselves as they start college; for example, students accustomed to interacting with individuals of similar background, religious orientation, and ethnicity might not know how they will behave when surrounded by a more diverse crowd; or, for that matter, how they might behave in the absence of parental authority. And the case obviously does not apply exclusively to college students: anyone thrust into a new and unfamiliar context might act in ways that are entirely unpredictable, because—to put it in the right terminology—the reasons to which one's reasons-responsive mechanisms are responding have changed radically.

Now we can widen the point further: encountering new situations is a standard fact of life. Whether one moves to a new country, finds a new job, begins a new relationship, or is forced to start shopping at a new supermarket, the mechanisms needed to deal with the situation will require adjusting. And, if we were to go even further, though this is not necessary to the argument, virtually *any* situation in which we find ourselves is, in some way, new. Of course we do not need to acquire new (or modify old) mechanisms to deal with *every* situation, just as we do not need to acquire new habits to cope with every new solicitation—most situations are similar enough to previous ones that we can seamlessly move forward with our lives. But if sufficiently new contexts *can* require modifications to the existing mechanisms, given our ignorance of how our mechanisms work (or might work in new contexts) it is unclear that a process of taking responsibility in childhood—or at any time—can make one responsible for any and all future actions. Rather, it seems like taking responsibility will have to be an ongoing process, not one that occurs before any action for which we can be held responsible.

So much for history. What about ownership? We can note from the outset that Fischer and Ravizza center their account on socialization. Kane already points out a key problem: one can be socialized in ways that are akin to brainwashing—he uses *Brave New World* as his fictional example—so that one fulfills all the conditions for taking responsibility, but does so on the basis of a responsibility-undermining sort of socialization. (Kane 2000) Although Fischer suggests that we can get around the problem, because there are ways (though perhaps not clear-cut ones) to distinguish brainwashing from "normal" socialization processes, a deeper problem lurks in the background. Since the entire account depends on a kind of social training, "it is by no means clear whether there is room in such a picture for a meaningful distinction between evidence-sensitive education and merely causally inducing indoctrination…. [The account] does not entail that the child is *learning to act reflectively*. He is simply being *assisted to internalize admirable values*." (Zimmerman 2002 223) So it looks like the entire process of taking responsibility is essentially a process of internalizing social norms, and this does not seem to give us enough for an account of ownership.

Fischer and Ravizza's account seems to work, roughly, like this: if I have taken responsibility, then—when I am held blameworthy—I have no reason to complain. Obviously people might refuse to take responsibility in particular situations, or they might—because they disagree with a particular social norm—refuse to accept the praise or blame as justified in a given case. But the point is that they lack grounds for complaint since they see themselves as fair targets, and at least understand that it is generally fair to hold them responsible for their actions. Thus the account seems to settle a problem: when is it *fair* to take certain reactive attitudes toward individuals? And their solution is this: it

is fair when the individuals can recognize that it is fair. But this isn't the problem of moral responsibility at all. The problem is figuring out when I *am* responsible, not when I can be held responsible without my putting up argument. And this seems to require that the mechanisms on which I act—and my endorsement of them—are *mine* in some sense that is deeper than the simple internalization of social norms.

I am suggesting, in other words, that there is some inconsistency between ownership and socialization in the semi-compatibilist account. The two are not mutually exclusive, of course. But internalization alone does not constitute ownership. To bring out the point more clearly, we can return to Heidegger's description of Da-sein's everyday being-in-the-world. As Heidegger informs us early on in *Being and Time*, Da-sein is characterized by its always-mineness [Jemenigkeit]. There is a sense in which Da-sein is always mine, and Heidegger accordingly designates Da-sein's being as *existence*, in contrast to presence-at-hand (also translated as "objective presence"), which is a mode that can characterize only beings unlike Da-sein. As characterized by *mineness*—which is essentially first-personal—Da-sein differs from all other entities, which are only third-personally accessible. And Da-sein's mineness means that, unlike the actuality that can characterize other entities, Da-sein is always characterized by its possibility. Heidegger lays out all these features—along with the themes that he will pursue throughout Division I and much of Division II of *Being and Time* immediately after his introduction of Jemeinigkeit:

> Da-sein is my own, to be always in this or that way. It has somehow always already decided in which way Da-sein is always my own. The being which is concerned in its being about its being is related to its being as its truest possibility… And because Da-sein is always essentially its possibility, it *can* 'choose' itself in its being, it can win itself, it can lose itself, or it can never and only 'apparently' win itself. It can only have lost itself and it can only have not

> yet gained itself because it is essentially possible as authentic, that is, it belongs to itself. (42-43)

There is a good deal to unpack here. Let me merely index some of the key features to be discussed later: First, Heidegger clearly connects the mineness of Da-sein with its essential possibility, authenticity. Second, because of this connection (to be explained later), Da-sein can win (or find) or lose itself, which means that mineness alone—in seemingly paradoxical fashion—does not guarantee actually grasping myself, but instead creates the possibility for *not* grasping myself. Finally, this relation between mineness and authenticity involves a choice: because Da-sein is *always* mine, its possibility is characterized by a decision ("it has always already decided") or choice concerning the way in which it is mine.

Here I want to address the key theme of Da-sein's losing itself, which will help illustrate the flaw in Fischer and Ravizza's view of responsibility. We can begin by asking a question immediately implied by the above: if Da-sein is always mine, and this mineness distinguishes Da-sein from other kinds of entities, what of the being of other people, other Da-seins? If there is an asymmetrical relation between my own being and the being of other entities in the world, does that mean that—at least from the perspective of my Da-sein—other Da-seins are essentially "mere things" for me, disclosed primarily within a referential context of use? The answer has to be no: if Da-sein is characterized by mineness, and this distinguishes it from all other entities,[59] this will be true of every Da-sein and not merely my own. But then what is needed is an explanation of how—given that I do not "see" the Da-sein of others *as* mine—I could distinguish them from

[59] I am leaving out here the difficult case of animals in Heidegger's philosophy, which places them in a sort of intermediate category between Da-sein and mere things, since animals on the one hand interact with other entities and use them, but nevertheless lack a world.

mere things. This account is important, since without it, it would seem as if other Da-seins *exist*, in contrast to mere things, and yet each Da-sein would systematically see other Da-seins only as things.

To see how other Da-seins are disclosed to us *as* Da-seins, we can return to Heidegger's account of our circumspect dealing with things. Every referential structure—every in-order-to that governs the use of tools—has its ultimate reference or in-order-to in Da-sein. But obviously this ultimate reference is not only one's *own* Da-sein. When I grade papers, I use paper, pen, and ink, a desk and a lamp, a chair and a reference book, and I use all of these in order to provide a grade for my student. In reading the paper, underlining strange word formations, adding comments and occasional question marks, I am not explicitly thinking about the student—I am focused on the content of the paper—but the student is the one for whom I do all of this, and that reference governs the care with which I read and comment and the content on my comments. I can perform the same activity for *any* of my students; I can even perform it for myself, when I revise a paper for publication. Similarly, a carpenter may build a chair *for* a client, or he may build it for himself. The activity will be the same, and so will the constitutive references governing it, though of course the particular identity of the individual for whom the activity is performed might change some details (I might, for example, make my comments to myself more cryptic than those I write for my students). Thus, all tool-using activity—all our agency, really—is ultimately performed for the sake of Da-sein. And it is performed not simply for *my* Da-sein—though it can be—but for any Da-sein. The point, in other words, is that activity is governed by a for-the-sake-of-which that determines the nature of the activity, and which is interchangeable and indefinite. And of course it has to be so:

otherwise every activity I perform for the sake of one individual would be radically different from an activity performed for another; but this is, again, not the case: my commenting on my paper and my commenting on a student's paper are the same kind of activity, even if the details vary.

Thus, not only does an indefinite Da-sein govern our activity as the recipient of its product, but the activity itself operates on the basis of publicly accessible norms. Not only do I use the pen and chair in order to return the paper to a Da-sein, but I use the pen and paper in a way prescribed *by* Da-sein. These two ways in which Da-sein's possibilities are social possibilities are, clearly, connected. On the one hand, Da-sein performs tasks for the sake of other Da-seins. Again, even if it performs the task for itself, it takes itself as one Da-sein among others. Since the Da-seins for whom the tasks are performed are interchangeable, the tasks themselves are publically accessible: I can brush my teeth and you can brush your teeth, and we will be performing the same activity; or, I can buy an ice cream for myself or for my friend, but I will be performing the same task. And since the tasks are publicly accessible, this means that the tools used in them must also be publicly accessible. Thus, other Da-seins are already disclosed within our use of tools as those for whom these tools are used: by walking down the street, I use the sidewalk as a tool. In using the sidewalk as a tool, what is disclosed to me is the nature of the sidewalk as something for the use of Da-sein as such—not my own Da-sein, but Da-sein in general. Heidegger refers to this as the representability of our being in the world with others. "In the everydayness of taking care of things, constant use of such representability is made in many ways. Any going to…, any fetching of…, is representable in the scope of the 'surrounding world' initially taken care of." (239) Since

both what we do and what we do it with are essentially public, anyone who does these things with these tools is essentially replaceable or representable by another.

We can see the significance of this if we recall that Da-sein first finds itself within its world. "This nearest and elemental way of Da-sein of being encountered in the world goes so far that even one's *own* Da-sein *initially* becomes 'discoverable' by *looking away* from its 'experiences' and the 'center of its actions' or by not yet 'seeing' them all. Da-sein initially finds 'itself' in *what* it does, needs, expects, has charge of, in the things at hand which it initially *takes care of* in the surrounding world." (119) As we have already seen, Da-sein is always thrown into a world, and it must find itself—if it is to do so—from out of that world. In Heidegger's account of Da-sein's sociality, or its essential being-with, the world in which Da-sein finds itself is already constituted by other Da-seins. Since the other Da-seins whose norms govern the world in which Da-sein initially finds itself are representable (this is, again, what makes the norms public, or capable of structuring the world of *any* Da-sein), Da-sein initially finds itself defined by a replaceable Da-sein, which Heidegger calls "the they."[60]

So here we have a problem. Da-sein is always mine. But this always-mineness does not guarantee that Da-sein has *found* itself in an authentic way. Instead, Heidegger muses, "what if the fact that Da-sein is so constituted that it is in each case mine, were the reason for the fact that Da-sein *is*, initially and for the most part, *not itself*?" (115-116) The question is rhetorical, as Heidegger indicates on the same page. And, as we have

[60] Of course the term Heidegger uses is "das Man," which has no workable noun equivalent in English. There are three options here: (1) One can simply write *Man*, which means the term cannot function as part of an English sentence. (2) One can use "one," as in "one can use…" (3) One can use "they," as in "they talk a lot, don't they" (*Pulp Fiction*). I will stick to "they," because I find it easiest to navigate grammatically. Occasionally, however, I will also use "one," and the context should make it clear that *das Man* is intended.

already seen, Da-sein's being always mine is precisely what allows it to win or *lose* itself. To see how this works, we need only consider the features already introduced: in its always-mineness, Da-sein's being matters to it. That is to say, Da-sein always projects and pushes ahead into its possibilities, and these possibilities are ones that essentially constitute it as the entity it is. Da-sein is as it understands itself. In its always-mineness, Da-sein takes its projected possibilities as its own. But the possibilities with which it initially finds itself saddled are ones understood in terms of its world and its taking care of that world. "'One *is*' what one does." (239) So the very feature of Da-sein that allows it to win itself, initially always leads it astray into taking the public possibilities in which it finds itself as its own.

Thus, Heidegger notes, "'the others' does not mean everybody else but me—those from whom the I distinguishes itself. They are, rather, those from whom one mostly does *not* distinguish oneself, those among whom one is, too." (118) Da-sein's self, in the everydayness in which it is initially and for the most part, "is the *they self* which we distinguish from the *authentic self*, the self which has explicitly grasped itself." (129) So in finding itself thrown into possibilities prescribed by the they, Da-sein does what one does, understands itself as one understands oneself, and is as one is. On the one hand, "the they itself, for the sake of which Da-sein is every day, articulates the referential context of significance" (129), which is to say that Da-sein's world is pre-given to it as meaningful in the way they see it as meaningful so that "the everyday possibilities of being of Da-sein are at the disposal of the whims of the others." (126) Da-sein's possibilities—it's for-the-sake-of-which—are given to it. But on the other hand, they are not given by anyone in particular, because "the others, as distinguishable and explicit,

disappear more and more. In this inconspicuousness and unascertainability, the they unfolds its true dictatorship." (126) Da-sein does not, in its everyday mode of being, notice that its possibilities are set out by others—the others withdraw in much the same way that tools withdraw in absorbed coping, so that Da-sein takes the self handed over to it by the they as its own. It does not see itself as something separate from others, and it does not see the norms, values, and self-understanding it has been given as something external or foreign to it; it accepts it at face value as its own, and this is the sense in which the self becomes a they self.

So what, exactly, is the result of this becoming a they self? The "dictatorship" of the they "prescribes what can and may be ventured," and thus gives rise to "the *leveling down* of all possibilities of being." (127) In other words—and this is the problem—it tends toward making possibilities actual, by giving them as fulfillable prescriptions instead of unfulfillable ways to be. Thus, it tends to lead Da-sein to see itself not as *existing*, in the technical sense, but as present-at-hand, as defined not by its possibilities, but by its actual properties. Of course Da-sein cannot literally become a present-at-hand entity, and it cannot literally trade in its possibilities for objective properties. In seeing itself as an objective thing, Da-sein is still *projecting* that self-understanding, and so it continues to exist as Da-sein; the point is only that, as a they self, it misunderstands itself. Heidegger brings out the extent of the misunderstanding in his discussion of Da-sein's falling prey.

Falling prey describes the way in which the they self distorts Da-sein's disclosure of the world constituted by attunement, understanding, and discourse. Heidegger describes three aspects of this phenomenon: idle talk, curiosity, and ambiguity. Idle talk

is a modification of discourse, which emphasizes communication over disclosing. Instead of bringing what is talked about to light in a genuine way, idle talk discusses it in "average" terms, terms that are understandable to everyone, giving the impression that the phenomenon discussed is also understood. Curiosity involves pursuing possibilities entirely for the sake of novelty, so that Da-sein does not "dwell" within any one possibility, but immediately abandons it to seek another. Just as in idle talk nothing is really disclosed but only the illusion of disclosure is given, so in curiosity possibilities are leveled down, so that instead of pressing forth into something genuinely new, Da-sein pursues trends and guesses what the future will bring, so that anything that happens is immediately "recognized" as something already foreseen. And ambiguity, finally, involves a fundamental confusion between what is genuinely understood and what is understood only in the shallow modes of idle talk and curiosity.

In all these phenomena, Heidegger notes, Da-sein becomes absorbed in the world and "lost in the publicness of the they. As an authentic potentiality for being a self, Da-sein has initially always already fallen away from itself and fallen prey to the 'world'". (175) So Da-sein, as its finds itself initially in the they, is already lost. But it is important to note that it hasn't simply gotten lost; it has lost itself, which it can do because it has the character of being always mine. That is, in attempting to grasp its mineness, Da-sein locates itself in the world; but by looking at the world, Da-sein is locating itself in the wrong place. To put it another way, falling prey is not simply something that happens to Da-sein; it is something that Da-sein does, and it does this because fallenness appeals directly to its concern with its own being. In seeking understanding, for example, Da-sein grasps the shallowness of idle talk. In seeking its own possibilities, it finds the constant

novelty of possibilities presented to it by the they. The public world provides Da-sein with a way—or rather innumerable ways—to satisfy its search for itself. Thus, Da-sein is "tempted" by the world and "tranquilized" into constantly feeling that it is gaining something important. But at the same time, it is "alienated" from itself and "entangles" itself so that it cannot see any possibilities beyond those offered by the they.

The they already lays out what Da-sein cares about, so that "the public way in which things have been interpreted has already decided upon even the possibilities of being attuned, that is, about the basic way in which Da-sein lets itself be affected by the world" (169-170), so that "we enjoy ourselves and have fun the way *they* enjoy themselves. We read, see, and judge literature and art the way *they* see and judge. But we also withdraw from the 'great mass' the way *they* withdraw, we find 'shocking' what *they* find shocking." (126-127) Some interpreters argue that Heidegger's notion of authenticity is simply identified with the first-person perspective as opposed to the second and third-person perspective we take on others. (Carman 2006, 2005) But this cannot be right: if the they modifies both our understanding *and* our attunement, it is clear that the they can enter into and structure our first-person perspective on ourselves; that is the point of referring to the fallen self as a they self. It *is* a self, with motivations, reasons for acting, values, desires, and so forth; and it is a self because it is characterized by mineness.

Consider a simple event: I am eating in a restaurant and, after the main course, it occurs to me that I want desert. It may well be true that I want desert only because that is what one eats at the end of a meal; but that does not change the fact that *I want* it. If I point to my desire as a justification for why I intend to order desert, I am providing a reason, and this may well be the reason on which I act. And what is interesting on

Heidegger's account is that the reason is *mine*. But, at the same time, it is not *my own*. Consider another example: I am walking on the sidewalk and not in the middle of the road. Why? To avoid the cars. That is a reason, a good one, and—in accordance with the McDowellian analysis given above—it may well be the reason on which I act. But at the same time I am walking on the sidewalk *because* that is what one does. This is not a reason; it provides no rational justification (though, of course, "this is what one does" *can* provide justifications and reasons for actions in some situations). But it is nevertheless an explanation of what I do. That I have a reason—"to get away from the cars"—does not make that reason my *own*; it is, after all, the reason one gives; it is an obvious explanation, barely worth mentioning.

It is a characteristic of the understanding laid out by the they that a justification seems obvious; so obvious that certainly one does not need to even have it explained. The explanation in its obviousness does not appear as something given to me. "The they is everywhere, but in such a way that it has always already stolen away when Da-sein presses for a decision. However, because the they presents every judgment and decision as its own, it takes the responsibility of Da-sein away from it." (127) And here we get to the heart of the problem: Da-sein's responsibility is "taken away" insofar as it exists as a they self. If my actions, along with all the reasons and motives prescribed for them, are already given to me by the they, by everyone and no one, then I am no more responsible than anyone else for my decisions or my values.

If this account of Da-sein's falling prey to the world according to possibilities laid out by the they is right, we can now crystallize a critique of all the approaches discussed so far. Fischer and Ravizza's account of taking responsibility, because it is grounded

entirely in a process of socialization, of internalizing the values and attitudes of others, explicitly lays out a notion of taking responsibility that involves a giving away of responsibility. That individual agents may be held responsible by others, and that they even hold themselves responsible, does not show that they *are* responsible; it could show only that they are participants in a discourse that misunderstands responsibility. But attributionism fares no better. The suggestion, cited above, that "in a sense, values cannot fail to be those of the agent who acts on them," simply refers us to the distinction between mineness and ownership. Yes, the values on which I act as a they self are *mine*; but they are not values I own. Simply laying out a framework on which our rationality—our will—is embodied in our actions and attitudes allows us to lay out the possibility of control, but control of the sort that can support agency, never mind responsibility, will need ownership.

But volitionism does not help either; as I argued in the last chapter, deliberation will not give us responsibility, because we would need to be responsible for the grounds of our deliberation as well. If the they already makes our decisions by handing us our values and motives, and if this is precisely how responsibility gets taken away, deliberation can only contribute to this process. Steven Crowell brings out this claim in his reading of Heidegger:

> deliberation takes place (as did the action from which it arose) within the constitutive rules of the 'world' in which I remain engaged. That is, I deliberate as that which I understand myself to be, in terms of my 'practical identity'… Thus, while only an individual can deliberate, I do not deliberate as my ownmost self. Rather, the reasons I adduce and the evidence that I find salient will normally be those typical of the current cultural, historical composition of the One… This does not mean that my reasoning is nothing but the rationalization of specific cultural conditions, but it does mean that the practice of deliberation, like all practices, is grounded ontologically in what is public, typical, and normative in a given community. That deliberation is explicitly oriented toward 'reasons for'

> does not, ontologically, get us any further than the analysis of everyday coping.[61] (2007a 53)

If my identity is taken over from the they, and if my deliberation necessarily takes place in light of my identity, then my deliberation will be no more expressive of my *own* self than my unreflective agency. Of course deliberation can contribute a great deal to make our actions more effective, or in getting them to better conform to social and moral norms—but none of this explains why deliberation would give us either ownership of responsibility.

But Fischer and Ravizza's account *does* add something. As I've already argued, *ownership* is precisely the condition we need for responsibility. I have only argued that their own view of ownership is insufficient. But it suggests two important points. First, explaining responsibility in terms of ownership, and especially in "subjectivist" terms, helps to shift the discourse concerning responsibility from the third-person perspective to the first-person perspective. So we should try to make sense of responsibility in terms of taking responsibility, which involves coming to *own* the mechanisms on which I act.

Second, Fischer and Ravizza see—correctly, I believe—that giving an account of responsibility will require distinguishing the temporality of action from the temporality of taking responsibility. As they insist, "it is necessary, in order for an individual to be morally responsible for his behavior, that a process of taking responsibility… has taken place at some point prior to the behavior." (1998 242-243) Where I think they go wrong is in the notion of priority involved. On their view, taking responsibility must be prior to action in a historical sense—a sense I earlier characterized as temporally shallow. That is, there is an event of taking responsibility (to be sure, it is a drawn out and complex event)

[61] The reference to the Dreyfusian term "coping" here strongly suggests that Crowell means to extend Dreyfus's view of absorbed coping not only to "mindless" activity, but also to deliberation itself.

that occurs at a particular time that is in turn earlier than the event of the action taking place. And this account, I have suggested, does not succeed precisely because it is temporally shallow: if every action is a new event, then taking responsibility—if it is to account for responsibility—would have to occur, if it occurred in time, at the same moment as the action, since every action takes place in new circumstances and no one can properly take responsibility for the way their mechanisms will act in new, unanticipated circumstances.[62]

In fact Heidegger's attempt to explain how Da-sein can be authentic given its initial falling prey appeals specifically to ownership and to temporality. It is to this account that I now turn.

[62] This is the extreme version of the criticism I gave above, where I conceded that not every action is new. But, first of all, enough actions *are* new that an earlier act of taking responsibility cannot cover them all. Second, especially in the case of unreflective actions where the agent simply responds to solicitations, it is always open to the agent to deny responsibility on the grounds that the solicitation in the present case was just different enough from past ones to throw off his nondeliberative mechanism. Consider here Sher's example of Father Poteet, who, though a competent driver, mistakenly believes he can weave seamlessly into traffic and ends up causing a massive accident. Just what is supposed to block the thought that Father Poteet saw a solicitation because the case was similar to previous ones, but that in fact the situation was slightly different from past ones, so that his normally reliable driving mechanism—the one he has taken responsibility for—failed to operate correctly? In fact, it seems like this is exactly what happened. But, had Father Poteet been aware that his mechanism *could* misfire in situations with differences imperceptible to him, he would perhaps not have taken responsibility for it as a nondeliberative mechanism, but would instead have been a less impulsive, defensive driver.

6

Temporality and Ownership

A. Discovering Ownership

I began the discussion of responsibility by examining the problem posed by Strawson's Basic Argument: that responsibility is impossible for a being that is not *causa sui*. This problem is already implicit in the earliest view of responsibility, in Aristotle's suggestion that we can get out of it by postulating a *joint responsibility*: perhaps we are not *entirely* self-created, but we can create ourselves just enough to make moral responsibility possible. This requires that we have the ability to produce something that is our *own*, and not simply given over to us by nature, by society, or whatever other external cause. Volitionism, which attempts to make deliberation a condition of ownership, falls short: our deliberation itself, for the most part, takes place against a background of values, ways of thinking, and attitudes that are not our own. Our thrownness into the world and our being-with the they prevents us initially from owning ourselves. Deliberation, no matter how explicit, cannot break free of our thrownness, because we are thrown as deliberating beings. Attributionism thus cuts out the deliberation and appeals directly to the idea of a self as defined by the coherence of its values and judgments. If a self is nothing more than such a coherence, and its actions follow from that coherence, then the self can be held responsible for those actions. But this does not help: if a self is *only* a coherence, then it is not clear why the actions it produces are any different from

the "actions" that result from the functioning of any other coherent entity, like a rock or a light bulb. Constitutivism, which will be a topic of this chapter, seemingly bypasses the need for absolute self-creation, but it must still give an account of the basis on which we might say that an entity is *responsible*: integrity and cohesion are not sufficient for ownership, because they are phenomena that occur in all sorts of entities that are not self-owned.

Finally, then, we reached the possibility of thinking of responsibility in temporal terms. One is responsible for an action if one has already taken responsibility for it, or come to own the character (or background) that produces the action, before the action has taken place. But if we take this "before" in terms of shallow temporality, we are no better off. The problem with volitionism and attributionism was that they could not separate what is one's own from what is not, and thus give an account of ownership as the ground of responsibility. But if taking responsibility is a merely historical process, then it is subject to the same charge. Fischer and Ravizza, of course, set out to argue that taking responsibility is consistent with causal determinism. But whether or not their argument succeeds, its reliance on socialization in taking responsibility means that the process is just as trapped in the they as any other. We are thrown into the possibilities laid out by the they. Taking responsibility, in Fischer and Ravizza's sense, only involves internalizing the possibility—given by the they—of taking others' judgments of our responsibility as appropriate. But *that* does not make them appropriate unless our taking responsibility can itself be our *own*. So the way out is to stop assuming that ownership must be understood as a historical process, that is, a process in which I start (as a child, perhaps) with no ownership and then proceed to take ownership of something.

Consider one of the most famous arguments in defense of ownership: John Locke's argument in *The Second Treatise of Civil Government*. (1980 18-30) Locke lays out the problem thus: originally, the earth and its products belong to all of humanity in common. How, then, can any individual rightfully come to own anything? Locke argues that there must be a solution, since each of us has a right to life, and thus a right to eat; and if I have a right to make something my own for the purpose of eating it, there must be a way of turning a common right of property into an individual right. Famously, the solution involves labor: since I own my body and the labor I perform with it, in laboring to turn the fruits of the earth into food, I mix something of my own with something that is owned in common. And through the mixing of what is my own, I make what I mix it with my own. Now clearly the argument has problems, and it is not my aim here to defend it. Let me just consider one problem: why would mixing something that is mine with something that is not mine make it mine? If I purchase a box of chocolates with a group of friends, the chocolates belong to all of us. They do not become mine alone simply by virtue of my performing the labor of fighting off the others and shoving the entire contents into my mouth! But Locke makes no such claim: I have no right to property that I cannot use. So my *right* to survive—for which I need the right to property—is constitutive of my ownership. My labor is co-constitutive with it. But in what sense is labor or my body my own? It is not my own in the sense of property, since it is not transferable. Even if I sell my labor or my body and it belongs to another as property, it still remains my own in the sense in which it was my own to start with—it is always mine.

So let me work out the key points. First, Locke does not ask how I can come to create ownership. He is asking how I can *transfer* ownership, from something owned in common to something owned singly (and, remember, if we ask where ownership comes from in the first place, the answers is pretty mysterious: God). I do so by mixing what is now owned in common with something that is always mine, that is, my labor (which, of course, is always mind in the sense that it is non-transferable, but also in the sense that it is not something I *came* to own at any point: as long as I have existed, it has always already been my own). And the resulting single ownership is legitimated by a teleological condition: my right to survive (which is also, in a sense, my own), for the sake of which I can take something owned in common and, mixing it with what is mine, make it my own. Taking this as a guide, I suggest the following features as constitutive of what is involved in genuine ownership: (1) current ownership, (2) always mineness, and (3) a teleological condition that enables 1 and 2 to produce genuine ownership. I propose reading Heidegger along these lines. I find myself now owned by the they.[63] The self owned by the they is already always mine. As a self characterized by mineness, my self "is essentially possible as authentic." (43) Key to this view is the recognition that I do not create myself wholly from nothing in order to become my own; rather, I take myself up *as* my own through a transfer of ownership in accord with my essential possibility.

This theme is central to Heidegger's early work in the form of authenticity [Eigentlichkeit], which—both in its linguistic root and in Heidegger's usage—might as

[63] Technically, the self is not initially "owned" in a genuine sense by the they. The they, as noted, is absolutely anonymous. So ownership by the they is not ownership strictly speaking, since it is ownership by no one in particular. Yet it is also not non-ownership. As Heidegger notes, "the they-self is an existentiell modification of the authentic self" (317), which suggests that it is something like a non-owning type of ownership; perhaps in the way that someone may forget that they own something, so that they remain the rightful user of the item but cannot make use of it.

well be translated as "self-ownership." The authentic self is distinguished from the inauthentic self, the self that is entangled in the world and the they. But Heidegger is always explicit that in taking ownership of ourselves we do not somehow step out of the world and the they—both are constitutive of Da-sein as care. This theme is one Heidegger maintains consistently. To take two examples:

> (1) "It is only on the basis of an *antecedent* "transposition" that we can, after all, come back to ourselves from the direction of things." (Heidegger 1982 161)
>
> In owning ourselves, we find ourselves out of a world through a transposition: we do not remove ourselves from the world, but rather first discover that we have already been looking in the wrong place.
>
> (2) "Understanding *can* turn primarily on the disclosedness of the world, that is, Da-sein can understand itself initially and for the most part in terms of the world. Or else understanding throws itself primarily into the for-the-sake-of-which, which means Da-sein exists as itself. Understanding is either authentic, originating from its own self as such, or else inauthentic. The "in" does not mean that Da-sein cuts itself off from itself and understands "only" the world. World belongs to its being a self as being-in-the-world." (146)
>
> The "transposition" of the previous quote involves a change of perspective: Da-sein can understand itself in terms of the world into which it is thrown, or it can understand itself in terms of its for-the-sake-of-which. In the latter case, Da-sein's understanding—which means, its projection in terms of which it is what it is—*originates* from its own self.

Obviously this is all schematic. The key is to understand how Da-sein is disclosed to itself as responsible. This happens in Heidegger's account of conscience, where Da-sein's responsibility is revealed. But then, "only in responsibility does the self first reveal itself—the self not in a general sense as knowledge of an ego in general but as in each case mine." (Heidegger 1982 137) In revealing the self as a responsible self, Da-sein can take over its ground: it takes its thrownness into itself in light of its ownmost potentiality of being, which is its anticipatory self-projection into itself existing as a whole. In other words, Da-sein compares itself to itself as a whole—a point I will try to make sense of—

and thereby discloses itself as an owned, free entity. In the last chapter, we left Da-sein scattered among the possibilities of being-in-the-world handed to it by the they. That is an image "of Da-sein as *fragmentary*." (233) Finding itself, for Da-sein, requires finding itself as a whole by projecting its ownmost possibility.

B. Constitutivism: Setting the Stage

In discussing free will, I suggested that Korsgaard's approach gives us a way—though an insufficient one—to avoid the difficulties that threaten both compatibilism and incompatibilism. I want to take her approach up again, though in a slightly different capacity. In developing the account I outlined earlier, Korsgaard has adopted a position that has come to be called constitutivism. Just as Korsgaard's account suggests a way out of the free will dilemma, I believe constitutivism can give us the strongest solution to the problem of responsibility. After briefly explaining the main points of constitutivism, I will attempt to explain why—properly carried out—it stands a chance of answering the Basic Argument. For contrast, I will then outline Frankfurt's position, which is both the starting point and the major target of constitutivism. I will then lay out the basics of Korsgaard's approach and suggests that it does not succeed because it implicitly relies on a constitutive aim that serves as a condition of possibility for the aim she proposes. In the following section, I will then argue that the needed constitutive aim is Heidegger's notion of authenticity, understood as anticipatory resoluteness.

Constitutivism is an attempt to explain how norms can be unconditionally binding. It functions like this: There are norms or aims constitutive of agency. Thus, in

order to be agents, we must be subject to those norms. But we cannot help but be agents; therefore, some norms—those that are constitutive of agency or that can be derived from other features constitutive of agency—are unconditionally binding on us.[64] Properly worked out, constitutivism stands to provide the strongest foundation for a theory of moral responsibility. The Basic Argument holds that the task of providing such a foundation cannot be met without the possibility of agents' being self-causing. But that is precisely what constitutivism provides.[65] Responsibility is responsibility relative to some norm. If the norms against which agents can properly be held responsible are, in fact, norms constitutive of agency, then agents, simply by being agents, necessarily create themselves in light of those norms. It is therefore reasonable to hold agents accountable in light of those norms. It remains true that agents cannot be responsible for the norms themselves, but this no longer undermines their responsibility: they are, as agents, responsible for the sorts of agents they are in light of the norms due to which they are agents in the first place. Agents are thus self-created in light of the very norms that make them agents. Another way of putting the point is this: acting on constitutive norms makes one's action autonomous; we often fail to act autonomously, of course, but we still act in

[64] For reasons of space, I will address only Korsgaard's account, but another major constitutivist approach is laid out by David Velleman, for which see especially Velleman (2000, 1992) and Chapter 3 of Velleman (2009). Very roughly, he argues that human beings naturally seek understanding, and note that there is a part of the universe—their own bodies—that they can control in such a way as to match their understanding. From a different starting point, he argues that in acting intentionally, we must know *why* we are acting. Both accounts suggest that, in order to act intentionally, we must guide our actions by our self-understanding, so self-understanding is constitutive of action (or, perhaps, only of full-bloodied action—Velleman's view on this point has changed over time).

[65] Though she spends little time on this point, Korsgaard says as much: "I think it is true that we could not rightly be held responsible unless we created ourselves, but false that that makes the idea of responsibility incoherent." (2009 130) Velleman does not draw this connection; when he comments on responsibility, it is only to make the traditional—and problematic—claim that even if what I do is not an "action" in the full sense, because it is not guided by self-understanding, I can be held responsible for not exercising greater self-control. (1992 465)

light of the norms that constitute autonomous action. Thus, we can be responsible for violating those norms.

As already noted, Frankfurt attempts to explain autonomy and responsibility by making use of the concept of reflective endorsement: we are autonomous, responsible, and free when we act on first-order desires with which we identify. Working out the notion of identification has been one of the key tasks of Frankfurt's work since "Freedom of the Will and the Concept of a Person." Here I will attempt a construction of one of his views. To say that a person identifies with a desire, for Frankfurt, is to say that acting on it furthers something he *cares* about. What we care about depends on what we love. And love is a volitional necessity: it places limits on what choices a person can make. (Frankfurt 1999b) Importantly, Frankfurt distinguishes between volitional necessities and other constraints on our choices. The unwilling addict, again, can serve as an example to make the point clear. The addict's desire for his drug places constraints on him: he cannot help wanting the drug, and perhaps he cannot help acting on that desire. But this desire is not one he identifies with, in the sense that pursuing the drug does not further (but, perhaps, harms) anything he loves. Thus, the desire is external to who he is; such desires "are generated and sustained from outside the will itself." (Frankfurt 2006 44)

Volitional necessities, on the other hand, are internal to the will. What makes them so is that the agent is not only constrained to act in accordance with them, but he *wills* to be so constrained. To love someone, Frankfurt argues, involves wanting to go on loving them. If I do not care whether or not I will still want to do philosophy in a year, then I do not really care about doing philosophy *now*. "Caring about something implies a diachronic coherence, which integrates the self across time." (Frankfurt 2006 19) This

means that what we care about determines who we are as persons; if I stop caring about something, or if I betray what I care about, there is a sense in which I am no longer the self that I was. In Frankfurt's famous example, Agamemnon is torn by a love for his army and a love for his daughter; in betraying one—killing his daughter—he violates his diachronic unity, destroying himself. (Frankfurt 1999a) So the self is constituted by its commitments: I am who I am because I care about the things that I care about, though most of these are contingent; a different person might not care about them at all. Frankfurt insists that this meaning of "person," on which a change in commitments signals the end of personhood, is a common one. Considering the case of a bully who has undergone a moral conversion, he points out that in extreme enough cases we might ordinarily say that he has become a new man. (Frankfurt 2002 124-125) My volitional necessities, then, determine who I am; if they change, I become a different person.

As the case of Agamemnon shows, however, things are not always simple: the things we care about might conflict. And this sort of conflict, which Frankfurt calls "ambiguity," is different from the case of the unwilling addict. The unwilling addict faces a compulsion from outside his will, so there is a truth about who he is. The ambiguous person faces a conflict from within, and thus there is no such truth. The only way out of ambiguity is wholeheartedness. When a person is wholehearted about his volitions and resolved with regard to one side or the other, his will has a definite shape: he is clear on what he cares about. When he is ambiguous, there is no truth about who he really is, "his will is in fact unformed. He is volitionally inchoate and indeterminate." (Frankfurt 1992 10) Becoming wholehearted thus involves giving one's will a reality by placing oneself firmly on one side or the other. Of course things are not always simple: one might solve

an apparent conflict, for example, by prioritizing the things one cares about. But one might also—like Agamemnon—be trapped in a tragic dilemma, where volitional necessities conflict directly and so the person cannot abandon either of them without ceasing to be who he is. Even more complicated, however, is the question of how one becomes wholehearted in the first place. Frankfurt has argued—controversially—that wholeheartedness requires satisfaction with one's self, that is, with one's volitional states. And this is not accomplished through anything the agent does or decides; satisfaction requires "no adoption of any cognitive, attitudinal, affective, or intentional stance. It does not require the performance of a particular act; and it also does not require any deliberate abstention. Satisfaction is a state of the entire psychic system—a state constituted just by the absence of any tendency or inclination to alter its condition." (Frankfurt 1992 13) To be wholehearted, then, an agent needs to have no tendency to change the state of his will; why he has no such tendency (perhaps he is simply tired, or has given up) doesn't matter. Frankfurt does add an important condition, however: satisfaction *must* involve the "entire psychic system." That is, the agent must not remain conflicted at all; and he must be satisfied not on the basis of repression or self-deception, but on the basis of a self-understanding. But, again, after that self-understanding is achieved, the agent either becomes wholehearted or he does not; he cannot *make* himself wholehearted.

In fact, it is *impossible* to become wholehearted through a decision or choice. Wholeheartedness, after all, is a matter of "really" being a particular person, and reality—even the reality of our will—is something independent of our will. We can, of course, try to resolve ourselves to wholeheartedness, but this does not guarantee that we will succeed; instead, we might simply cover up our ambiguity and simulate

wholeheartedness, which will fail to hold up when we are faced with a conflict. To be free, autonomous, and responsible, then, is to be wholehearted. To be wholehearted is to be satisfied with who we are. But we lack volitional control over whether or not we *are* satisfied. Frankfurt appeals here to Spinoza's notion that the highest good is "acquiescentia in se ipso." Rejecting the usual translations ("self-contentment" and "self-esteem"), he writes:

> There is something to be said for a bluntly literal construction of his Latin. That would have Spinoza mean that the highest good we can hope for consists in *acquiescence to oneself*—that is, in acquiescence to being the person that one is, perhaps not enthusiastically but nonetheless with a willing acceptance of the motives and dispositions by which one is moved in what one does… When we are acquiescing to ourselves, or willing freely, there is no conflict within the structure of our motivations and desires… The unity of our self has been restored. (Frankfurt 2006 17-18)

One attains freedom and becomes responsible for what one does, then, when one is wholehearted and does not act contrary to the demands of one's volitional necessities. Let me reiterate a key point: for Frankfurt, we do not constitute ourselves. Our selves are constituted by our volitional necessities, by the things we care about. But we do not control what things we care about. Nor do we control whether we are coherent persons at all: this results only from acquiescing to oneself. Acquiescing to oneself, however, means something interesting. It means finding oneself. But finding oneself in such a way that, until we have done so, we *have* no self.

Korsgaard takes up the themes of reflective endorsement and identification, but she rejects the premise that the self is something we *find*. Rather, we constitute ourselves in acting, and in fact we *must* do so, with the result being that self-constitution is the constitutive aim of agency. We have already seen much of this account, so I will be brief. Korsgaard argues that any decision—any choice of an action—involves two constitutive

norms, efficacy and autonomy.[66] To decide on an action, we must decide to succeed (for example, one cannot decide to turn on a light but refuse to flip the switch), and we must decide to make ourselves the cause of the action; that is, the action must be self-determined. Korsgaard argues that, as rational agents with reflective distance from our desires, we are under the necessity of *choosing* whether or not to act on them. Agency is inescapable for us: "It is our *plight*: the simple inexorable fact of the human condition." (Korsgaard 2009 2) But, as Korsgaard tells us again and again, acting necessarily involves an agent: there is a difference between an event being merely caused by a desire, and its being caused by an agent. The former is not an action at all. As the aforementioned argument against particularistic willing was supposed to show, we cannot will non-universally, since then there is no difference between an action's being caused by an agent, and it's being caused by a desire. So in *choosing* to act, we necessarily commit ourselves to a universal principle, i.e., a principle that extends beyond the situation at hand. And in committing ourselves to such a principle we *identify* with it, that is, we choose it as constituting our will. Thus, Korsgaard can argue that Frankfurt's account has things backwards, at least insofar as it has autonomous actions being ones that arise out of a self that the agent has discovered as his own. "The intimate connection between person and action does not rest in the fact that action is caused by the most essential part of the person, but rather in the fact that the most essential part of the person is *constituted* by her actions." (Korsgaard 2009 100) So any time we decide on a

[66] She identifies these with Kant's hypothetical imperative and categorical imperative, respectively. (Korsgaard 2008b 82-83) I will not take up here the question of whether her interpretation of Kant has much to do with Kant or, for that matter, whether there is in Kant anything like "the Hypothetical Imperative."

particular action, we must thereby commit ourselves to having a particular kind of will, or being particular kinds of persons: self-constitution is the constitutive aim of agency.

A key difficulty immediately manifests itself. Korsgaard argues that something only counts as an *action* if it succeeds in constituting the agent. But then it seems we lose any notion of normativity. Normativity, per general agreement, involves the possibility of getting things wrong: an unbreakable standard isn't a norm at all. But if we fail in our constitutive aim, then it seems we have not acted at all, so violating the norm appears impossible. (Lavin 2004) Korsgaard's reply relies on a bit of Aristotelianism: a good harp player and a bad harp player are not performing different activities; they are performing the same activity, but only the former is performing it excellently. Similarly, Korsgaard relies on an analogy that recurs throughout her work: to build a house, one must follow certain standards. "A good house is a house that has the features that enable it to serve as a habitable shelter—the corners are properly sealed, the roof is waterproof and tight, the rooms are tall enough to stand up in." (Korsgaard 2008b 112) These are internal standards constitutive of something's being a house. Something that deviates too far from the internal standards is not just a bad house; it isn't a house at all. So "even the most venal and shoddy builder must try to build a good house, for the simple reason that there is no other way to try to build a house." (2008b 112-113) An obvious objection is that of course someone could try to build a bad house by saving money on everything, using the worst materials and the cheapest paid labor, and so on, so that the result has a leaky roof, walls that can barely survive in high wind, etc. But Korsgaard sensibly points out that such a person is no longer involved in trying to build a house: he is trying to build a simulation of a house that is good enough to fool someone into buying it. Deviating too

far from the standards constitutive of house-building involves participating in a different activity. Similarly, a bad action—one that fails to be autonomous—does not on that count cease to be an action. "Obviously, it doesn't follow that every action is a good action. It does, however, follow that performing bad actions is not a different activity from performing good ones. *It is the same activity, badly done*." (2008b 113)

So just how is defective action possible? Korsgaard suggests that agents can, in fact, simply follow their desires. What they *cannot* do is shrug off the norms constitutive of agency. Through a reading of Plato's *Republic*, Korsgaard argues that it is possible for agents to act on principles that fail to effectively constitute them in the following sense: an agent might follow a principle of prudence, or a principle of honor, or simply a principle of following whatever desire he happens to have. All of these serve to unify, or constitute, the agent under a single principle. But the principle, if it succeeds in creating stability in the agent, does so only contingently. The clearest example is that of choosing to act on whatever desire one has at the moment. If my desires keep changing, it will be impossible for me to get anything done, since each new desire will distract me from whatever I was doing on the basis of the previous one. So if I *do* manage to get anything done, it will be the result of a lucky accident, because I will be leaving it entirely up to chance whether or not new desires distract me from what I am doing. And action on such a principle is defective, because it fails to genuinely unify me as an agent. But—tying her account to Kantian universalizability—Korsgaard notes that "it is only when you ask whether your maxim can be a universal law that you exercise the self-conscious causality, the autonomy, that yields an action that can be attributed to you as a whole person."

(Korsgaard 2008b 124) So a defective action is one that fails to be autonomous, and thus *cannot* be attributed to me as a whole person.

All actions, including defective ones, *are* attributable to us, however; otherwise they would not be actions. "An action is yours when it is chosen in accordance with your constitution. Your constitution is what gives you the kind of volitional unity you need to be the author of your actions." (2008b 125) What makes an action *mine*, what establishes ownership, is my constitution. But this constitution arises in the action itself. This, of course, is why Korsgaard can argue that if an action is *too* defective, deviates *too* far from autonomy, is ceases to be an action at all: it fails to constitute me as unified in any sense, and so there is no one acting. But a defective action that is still an action does constitute me as a unified self, though it does so poorly, and fails to fully unify me because it fails to guarantee diachronic unity. But who is it that fails to be fully unified? In order for me to be able to recognize that my willing has failed, I must be someone—in Korsgaard's terms—over and above the self constituted in the defective willing. Otherwise, whether I will autonomously or not would itself be an accident. And this would return us to something like Frankfurt's account, where the self is constituted in wholehearted autonomy, which is itself entirely out of our control. Here we are back to the problem noted in my previous discussion of Korsgaard: the self-constitution involved in each act of willing seems to already presuppose an underlying will on the basis of which self-constitution is possible as such. Something must unify my defective and autonomous willing in order for both of these to be my willings, and in order for the latter to itself be produced by something other than a defective, accidental mode.

The problem is similar to another one: if the self is only constituted in its willings, this seems to imply that it does not exist *prior* to willing, and we must then ask who is undertaking the willings in the first place. Korsgaard argues that we become selves by choosing in light of—and thus endorsing—principles that make up our specific practical identities. In so doing, we make those identities our *own*. It might thus seem as if we must already have an identity in order to endorse an identity, since we must have a standpoint in light of which some reasons are salient. But Korsgaard argues that this objection rests on a misunderstanding, since "it assumes that the endorsement of our identities, our self-constitution, is a state rather than an activity." (2009 43) Korsgaard compares this to the case of an animal that constitutes itself—continuing to exist as the animal it is—by following its principles, which are its instincts. No one asks how a giraffe can constitute itself if it must already exist prior to following its instincts. And agential self-constitution works the same way: it is an activity, or process, that is ongoing throughout the course of an individual's life. This may be right, but Korsgaard's account doesn't explain how it is possible. To see this, we can look once again at the problem of pre-deliberative agency.

Much of what we do is non-deliberative. Korsgaard, of course, recognizes this fact, and is willing to accept that non-deliberative activity can still count as action, and not simply a reversion to animality. It can still involve principles, albeit not explicit ones. "Acting on a rational principle need not involve any step-by-step process of reasoning, for when a principle is deeply internalized we may simply *recognize* the case as one falling under the principle." (Korsgaard 2009 107) But this seems to make non-deliberative activity dependent on self-consciously chosen principles via tracing. And that suggests that pre-deliberative agency—that is, non-deliberative agency occurring in

the absence of tracing—is impossible. This means that acting on anything other than a self-chosen principle is impossible, and that already seems highly suspect. And this further raises the question of how principles dependent on self-consciousness *can* become "internalized" in the first place: once a principle is internalized in this sense, self-consciousness is absent from it. And thus *I* am not involved in applying the principle to particular cases. (Crowell 2007b) What Korsgaard *cannot* account for is complete self-creation, which is precisely what she is aiming to do. But we already find ourselves with certain principles, principles—certain ways of responding to our surroundings—that afford actions, and yet which we do not self-consciously choose; of which we are not, in fact, conscious. This should be clear from the prior discussion of deliberation: what we consider in deliberating is not up to us; it springs from who we already are. And even when we do deliberate, "we need a way to distinguish deliberation directed by the agent from reasoning processes in the agent that mimic such deliberation but are not directed or endorsed by the agent." (Bratman 2001 317) This, of course, is what the constitutive account is trying to do, but since our deliberative mechanisms spring from our prior identities, and must ultimately spring from pre-deliberative mechanisms, self-constitution would only account for *ownership* of the relevant deliberative processes if it could do so retroactively. There is no hope of fully identifying the self that acts with the self that chooses through CTD. In order to be able to constitute ourselves in Korsgaard's sense, then, we must already have a way of retrieving ourselves, or finding ourselves in the possibilities we have been thrown into. We need a constitutive principle that allows for ownership of non-deliberative selfhood.[67]

[67] Again, though I cannot go into it in detail, I think there is an argument to be made for a similar response to Velleman. Unlike Korsgaard, he does not think we can constitute our *selves*. The self is only a reflexive

I concluding, I want to draw attention to several points. First, despite Arpaly's claim to a "whole self" view, attributionism has no monopoly on that notion. Frankfurt's talk of satisfaction as involving the entire psychic system, his discussion of wholeheartedness, and his account of volitional necessities are precisely meant to delimit the boundaries of the self. Korsgaard is also clearly after a notion of wholeness; for her, the agent as a whole is supposed to be constituted in a decision. Second, ownership is central to all accounts, which hold it as a key component in freedom or, at least, autonomy. Finally, as I have been suggesting, we come closer to a solution by finding a constitutive aim that underlies and makes possible the agential aims of Korsgaard's account. Although Korsgaard lays claim to an account of the constitutive aim of *agency*, I have been suggesting that her account falls short of giving us agency as a whole. It gives us only norms for *actions*, or for choosing actions. What I will now look for is an aim that constitutes agency as a whole, which will involve finding oneself, thus moving (in a sense) closer to Frankfurt's account. I will argue that this aim is anticipatory resoluteness.

C. Conscience and Resoluteness

As noted, we left Da-sein at the end of the last chapter in a fragmentary state, scattered among innerworldly beings, entangled in the world, and fully in the thrall of the they as a they self. Removing all of these obstacles requires finding a way to *individuate* Da-sein, that is, to find a way to free it from the they and from innerworldly beings.

---

perspective. (Velleman 2002) But Velleman does think self-understanding must constitute agency. However, since agency can also constitute self-understanding—since we do act in ways we do not (yet) understand—his aim presupposes that agency and understanding have already been brought together. And this is what Heideggerian authenticity provides.

Heidegger lays out this possibility in what he calls the fundamental mood of *Angst*. Unlike fear, which is always fear of something in the world, *Angst* is not about anything in particular. What it is about is not something *in* the world, but about "being-in-the-world as such." (186) Like every mood, *Angst* discloses, but what it discloses is the world without significance. Since *Angst* discloses the world, it still discloses it as a referential totality, but in such a way that "innerworldly beings in themselves are so completely unimportant that, on the basis of this *insignificance* of what is innerworldly, the world is all that obtrudes in its worldliness." (187) The world obtrudes, or is experienced as a burden, because in *Angst* Da-sein doesn't know what to do with it: it is faced with a familiar referential framework, but one that lacks any solicitations. It presents no reason to do or want anything. In opposition to the tranquilizing falling prey, which draws Da-sein in through the semblance of complete understanding and perpetual seeking after possibilities, *Angst* discloses a world in which Da-sein is not at home, an *uncanny* world. What makes Da-sein at home in the world is its existence, it's understanding of itself as its being-in-the-world. Uncanniness is thus the term for Da-sein's recognition that it's "fit" with the world, it's being at home in it, is dependent on its projection of possibilities.

So *Angst*, in preventing Da-sein from finding significance, discloses Da-sein to itself apart from its factical involvements with entities and their usual importance. By thus separating Da-sein from the significances bestowed on the world by the they, it "individuates Da-sein to its ownmost being-in-the-world which, as understanding, projects itself essentially upon possibilities. Thus along with that for which it is anxious, *Angst* discloses Da-sein as *being-possible*, and indeed as what can be individualized in

individuation of its own accord." (187-188) In other words, since in *Angst* Da-sein does not project itself onto concrete possibilities, and in fact *cannot* so project itself, what is disclosed to it is simply itself as projection, without those possibilities being owned by the they and without being entangled in them but, of course, also (and by virtue of) without being able to press into those possibilities. To see how this works, we can take up Blattner's attempt to differentiate between thin and thick senses of existence. In the thin sense, Da-sein is concerned about its own being; in the thick sense, it presses forward into particular factical possibilities, thus filling out that being. (Blattner 1994, 2006) In *Angst*, Da-sein is still concerned about its being, and so is still seeking to press forward into possibilities, but it cannot do so because all possibilities have been stripped of significance.

This means that in *Angst*, Da-sein discovers its "true self," but in an empty sense: its true self is just being-possible, or understandingly projecting itself upon possibilities. What makes these possibilities non-empty is whatever we normally understand ourselves as, which defines who we are. Our self-understanding—our being anything at all in particular—is what is stripped away, leaving only the bare structure. *Angst* thus "reveals in Da-sein its *being toward* its ownmost potentiality of being, that is, *being free for* the freedom of choosing and grasping itself. *Angst* brings Da-sein *before its being free for… (propensio in)*, the authenticity of its being as possibility which it always already is." (188) Since Da-sein is disclosed as possibility, but not as any concrete possibility, it recognizes itself as free to choose itself. Of course being free to choose, or grasp oneself also opens the possibility of losing oneself, so *Angst* discloses the possibility of authenticity *and* inauthenticity. But both possibilities are now disclosed as possible on the

ground of choice. Rather than having to take the world as it is given, Da-sein sees its freedom to project its own possibilities. This does not give Da-sein any absolute freedom—it is still always in a world. But it serves to individualize it, and to disclose itself as possibility, though without any content. This means that *Angst* by itself discloses Da-sein, but does not allow it to *be* itself. It is only a preliminary to authenticity. Recognizing one's freedom is, we might say, a preliminary to using that freedom.

Authenticity enters the picture when *Angst* is expressed in discourse, in the "voice of conscience." Conscience is a call from Da-sein in *Angst*, in its uncanniness, to its self lost in the they. And what conscience reveals to Da-sein is its guilt. This guilt discloses a double "nullity" within Da-sein. First, as thrown, Da-sein has not thrown itself, nor has it, so to speak, prepared the pillows for its throw. After all, Da-sein is characterized by its facticity, and it always discovers itself among particular possibilities laid out for it by the they. To put it in standard English: we do not choose the world we are born into, nor do we choose the ways of life that world presents to us as options. But as we have seen, Da-sein exists "only by projecting itself upon the possibilities into which it is thrown." (284) Thus, Da-sein exists *as* something, and must always exist as something, but it does not give itself the "as-what." In a straightforward sense (though not an exclusive sense) Da-sein cannot be *causa sui*. "The self, which as such has to lay the ground of itself, can *never* gain power over that ground, and yet it has to take over being the ground in existing. Being its own thrown ground is the potentiality-of-being about which care is concerned." (284) Here we immediately see a crucial point. As Heidegger reminds us over and over, he does not—despite the consistently negative characterization of guilt as a "not" and a "nullity"—take guilt to be a *lack*. To lack something is to have something

missing—for example, if I am subject to a law and fail to follow it, this is a lack on my part. But Heidegger rejects this idea as a determination of Da-sein's existence, because the notion of lack belongs to entities unlike Da-sein. (283) An apple, for example, can lack something when we take a bite out of it; a train with a broken engine lacks functionality.

Da-sein does not lack something in the way an apple or a broken train does: being guilty is a *positive* characteristic of Da-sein. We already see a glimpse of that in the above quote, a point I will elaborate later: Da-sein is a *not* in the sense that it is characterized by an inability to "gain power over" its ground. But this inability is a positive feature of existence: Da-sein must exist *as* this ground. Thus, the inability characterizes an ability: Da-sein's being-possible means that it always is its possibility, and can come to *own* its possibility. More importantly, it can do so only *because* it is fundamentally guilty. Consider a being of infinite power, equipped (somehow) with the power to be entirely *causa sui*. Such a being could never *own* any possibility: since it has absolute power over all alternatives open to it, it can never really *be* its possibility, since it could always simply flit from possibility to possibility. None of its possibilities could define it, and thus none could be its own. It is at least partly in recognition of this that much of theology tends to characterize God as eternal rather temporally self-created and as having some inabilities, since (for example) God cannot act against what is best or diminish His perfection.[68] Unlike God, we are not perfect. But like God, we can *be* something, and this is possible only on the ground of guilt.

[68] For a clear argument to this effect, see Ch. 7 of the *Proslogion* in Anselm (1995). The comparison is, of course, flawed, since God is supposedly actuality—though not in the mode of an innerworldly entity—and not possibility. This is why God has to be eternal—he must always exist as actuality. So our guilt clearly

But Da-sein is guilty in a second way as well. As existing, Da-sein must live within some factical possibility. It "stands in one possibility or another; it is constantly *not* other possibilities and has relinquished them in its existentiell project." (285) And here, in an odd turn, lies Da-sein's freedom. "Freedom *is* only in the choice of the one, that is, in bearing the fact of not having chosen and not being able to choose the others." (285) Da-sein not only exists among possibilities that it has not itself made, but it must necessarily choose from among these possibilities. Its freedom is precisely this necessity of choosing. Da-sein is not free in the sense that is *could have* chosen some other possibility. This absolute sense of freedom is absent here and is quite possibly attributable to the they, since only the they allows Da-sein to pursue constantly new—though illusory—possibilities all at once. Rather, Da-sein's freedom is disclosed to it in guilt because other possibilities do appear to it *as* possibilities, and furthermore as possibilities that have *not* been chosen. It is in and through this disclosure that Da-sein can come to *own* its possibility: it comes to own it precisely as the possibility that it *has always already* chosen.

Again, we see a positive characterization in guilt: the possibility that Da-sein *is* is revealed to it as something it has always already chosen. As in the first disclosure of guilt, this second dimension allows for ownership: an entity that can be all possibilities at once, that does not exclude any possibility or way of being by choosing another possibility, cannot genuinely own its being. To be everything is to have no character. Unlike innerworldly things, which have a fixed essence, Da-sein is always possibility, so that *its* essence is never fixed. But it *is* that possibility, rather than all possibilities, and

---

sets us apart from God; the point, however, remains: an entity that created itself in time could not own itself.

thus it can be concerned with its being. Furthermore, it should be clear that it is on the basis of this guilt that Da-sein can first be responsible in the ordinary sense: to be responsible in the ordinary sense is to be subject to a norm, which one may follow or violate, and to *own* that adherence or violation, or to have it attributable to oneself *as* a self. That is precisely what guilt, in this second sense, allows. Someone who *can* be all possibilities at once cannot be genuinely subject to a norm, since one can deviate *and* adhere at once. Of course such an entity could—in theory—choose only to deviate or adhere. But someone who *can* deviate and adhere at the same time is not genuinely *subject* to the requirement to *be* a self. No doubt this seems question-begging: if I can be both A *and* B, but choose to be A, doesn't that make A *more* my own? But the contrast here is not between being able to be both A and B, on the one hand, and being able only to be A, on the other. Rather, it is between being able to be both A and B, or being able to be *either* A or B. Only the latter is placed under the necessity of having the choose himself; his *being* something is demanded of him. The former can indeed be A or B, but is under no requirement to be so; his being something is not *constitutive* of his self, but only an expression of that self.

But in what sense has Da-sein "always already chosen" itself? Aren't we all born into a particular social group? And doesn't that group map out for us at least our initial possibilities, those in terms of which we define ourselves even before being able to reflect on our options? This is slightly beside the point, in a sense: Heidegger grants—in fact, insists—that falling prey belongs to the care structure. This is why Da-sein *initially* and for the most part exists as a they self. Therein lies its guilt: in hearing the call of conscience, Da-sein "must bring itself back to itself from its lostness in the they, and this

means that it is *guilty*." (287) And indeed, in apparent contradiction to my reading, Heidegger does claim that Da-sein becomes authentic "by *making up for not choosing*." (168) But it turns out that Da-sein does not choose only in the sense that "the they even conceals the way it has silently disburdened Da-sein of the explicit *choice* of these possibilities" (268) and it "has let itself be given such possibilities as are prescribed by its public interpretedness." (270) This suggests that Da-sein must make up for "not choosing" only in the sense that its having chosen is hidden from it, so that it must reclaim or retrieve that choice in authenticity. Once again, when Da-sein *can* find itself, it finds itself as having *always already* chosen. And this is what is revealed to it in the call of conscience. Protesting that, after all, it had no choice is simply a refusal to heed the call.

But this still seems fishy. Did Da-sein *really* choose its possibility? This question rests on the idea that Da-sein *could have* chosen a different possibility. But this *could have* is not disclosed in guilt. The question of whether or not this choice could have been made differently is both unanswerable and irrelevant to Da-sein. Da-sein is disclosed in guilt as, first, having before it a field of possibilities which it has not created itself and from among which it must choose. And it is disclosed in the second place as already having chosen, and chosen in such a way that it has *not* chosen other possibilities. Da-sein owns its possibility because it has chosen it, not because it *could have* chosen otherwise. There is here no further question to ask about whether or not Da-sein has *really* chosen. This question adds nothing at all; it cannot be answered, because there is no possibility of *really* choosing or *not really* choosing. Since Da-sein is disclosed as having chosen a possibility and as not having chosen others, the question of whether it

ever *really* chooses is ontologically meaningless. In other words, Da-sein is disclosed to itself, through conscience, as guilty and therefore as fundamentally capable of self-determination in the sense of owning its possibility. To fail to be self-determining in this sense is not to somehow fail to choose, but rather to "forget," under the influence of the they, that one has already chosen.

Of course this point is controversial, and most Heidegger readers avoid it. Crowell, for example, in an impressive reading of the account of conscience, suggests that in revealing its grounds as beyond its power, conscience thereby reveals to Da-sein the ability to choose—and this first allows Da-sein to take up its grounds *as reasons*, that is, to act in light of norms and not merely according to them. (Crowell 2007a, 2008) This suggests that Da-sein's choice is always only in its listening to conscience; it does not discover that it *has* already chosen, but only that it *can* choose. This reading thus wipes out the temporal account I am giving, and also brings Heidegger closer to common sense. On my reading, conscience discloses Da-sein's deep temporality (as I have been calling it): we have not chosen our possibilities at some previous point in time, but rather we encounter them in our thrownness as something that we *have already* chosen. And that is puzzling, to say the least. The above objections, thus, make a good deal of sense, and I will have to postpone a fuller defense until the discussion of temporality in the next section.

Conscience "calls back by calling forth: *forth* to the possibility of taking over in existence the thrown being that it is, *back* to thrownness in order to understand it as the null ground that it has to take up into existence." (287) So guilt places an inescapable demand on Da-sein: to project itself onto the possibilities into which it has been thrown,

to exist as its ground. Furthermore, it is important that this demand—while it comes from Da-sein in *Angst*—does not reach Da-sein in *Angst*; the Da-sein in *Angst* cannot project itself onto its possibilities, but the Da-sein who hears the call can. At the same time, this Da-sein is summoned "to one's *own self*. Not to what Da-sein is, can do, and takes care of in everyday being-with-one-another, not even to what has moved it, what it has pledged itself to, what it has let itself be involved with. Understood in a worldly way for others and for itself, Da-sein is *passed over in this call*." (273) Thus the self reached by the call—beyond its worldly commitments—is deeper than the self as constituted by its volitional necessities, or the self constituted by its reasons. But what does this Heidegger-speak mean? Again, a fuller discussion has to wait until the next section, but the suggestion seems to be that *Angst* is always, as a threat, constitutive of Da-sein as care, though of course it is always possible to ignore it and go one with one's commitments, volitional necessities, and endorsements. The alternative is to hear the call and be "summoned" into one's own self—that is, ownership of one's possibilities.

Key to the structure of the call of conscience is that it is silent—unlike the "idle chatter" of the they, conscience speaks without words. And so the proper understanding response to it is, likewise, not verbal; it is reticent. But the point of the reticence is not simply that Da-sein, in understanding the call, does not speak—it responds by being *resolute*, resoluteness being "the *reticent projecting oneself upon one's ownmost being-guilty which is ready for* Angst." (279) But resoluteness is not "passive," since projecting onto being guilty involves being summoned to take over the ground. So in reticence, Da-sein responds by acting: "Understanding the call, Da-sein lets its ownmost self *take action in itself* in terms of its chosen potentiality-of-being. Only in this way can it *be*

responsible." (288) It is worth noting that "responsible" here translates "Verantwortlich," "answerable." This distinguishes it from the previous references to responsibility in the discussion of conscience, which use "Schuldigkeit."[69] This transition is significant: in *heeding* the call, which reveals Da-sein's guilt (Schuld), Da-sein takes action and becomes answerable for itself: in other words, it takes responsibility for itself.

On the other hand, Da-sein does not, as in the constitutivist and volitional theories, take responsibility for its action in the standard sense; despite frequently using the term, Heidegger suddenly tells us that "resolute, Da-sein is already *acting*. We are purposely avoiding the term 'action.' For in the first place, it would have to be so broadly conceived that activity also encompasses the passivity of resistance. In the second place, that term suggests a misinterpretation of the ontology of Da-sein as if resoluteness were a special mode of behavior of the practical faculty as opposed to the theoretical one." (300) The latter point signifies that resoluteness is both acting and understanding; it is disclosive agency, we might say, since in letting Da-sein choose—and own—itself, it both discloses Da-sein to itself and does so through Da-sein's agency in the world. But resoluteness does not mean taking action in the usual sense; Heidegger clearly means agency as such—not, that is, particular acts, but the projecting of possibilities that structures those acts or omissions. Resolute Da-sein is a self that owns itself. Its agency, thus, is not scattered among the they and the many things to be taken care of in the world. This is not to say that resoluteness is removed from the world, of course, as Heidegger stresses that this is impossible (world is, after all, constitutive of Da-sein). Rather, resoluteness re-enters the world as self-owning, disclosing the "situation," a term I will take up later.

[69] Macquerrie and Robinson's translation notes the distinction; Stambaugh's, unfortunately, does not.

Finally, conscience does not tell Da-sein how to act; it gives no concrete guidance at all. It cannot, as Heidegger notes, since it bypasses Da-sein's commitments to taking care of things; it "passes over *what* Da-sein understands itself *as* initially and for the most part in its interpretation in terms of taking care of things." (273) So the call is silent and not action-guiding; otherwise, "with its unequivocally calculable maxims that one is led to expect, conscience would deny to existence nothing less than the *possibility of acting*." (294) That is, if conscience *were* action guiding, it would *block* Da-sein's ability to act. This may seem odd, but the point is fairly clear: since Da-sein exists as possibility, the imperative it receives from conscience is simply to project its ground upon its ownmost possibility. But just what this consists of has to be determined by each Da-sein itself; truly acting, for Da-sein, involves being-possible. Thus, in fully acting as itself, Da-sein still exists *as* possibility. But to give it definite criteria for action would be to define an actuality for which Da-sein, as possibility, must strive. Falling prey in the they, among its guidelines and rules, prevents Da-sein from truly acting; it even reduces genuine *willing* to mere *wishing* (194-195), since all possibilities for action are already pre-given, so that Da-sein need not do anything but conform and wait for results rather than accomplishing them. If conscience were to lay down rules, it would simply reduce Da-sein back to the level of the they.

Consequently, the response to conscience—resoluteness—also cannot be a matter of following any pre-given rules. "But to what does Da-sein resolve itself in resoluteness? On what is it to resolve? *Only* the resolution itself can answer this. It would be a complete misunderstanding of the phenomenon of resoluteness if one were to believe that it is simply a matter of receptively taking up possibilities presented and suggested. *Resolution*

is precisely the disclosive projection and determination of the actual factical possibility." (298) That is, resoluteness involves genuine acting, which itself discloses the situation that calls for action. It does so by projecting onto Da-sein's being guilty, that is, by taking up Da-sein's ground and existing out of it. And since each Da-sein exists as being-possible, there cannot be any possibility that, in advance, all Da-sein must take up. At least, there cannot be any such *public* possibility; there is one possibility that is for every Da-sein its ownmost possibility: death. And it is here that I now turn.

D. Death and Being-a-Whole

Remember that at the end of the last chapter, we had left Da-sein in a "fragmentary" state. It is fragmentary in two senses. First, as absorbed in the they, Da-sein is scattered among the entities that it takes care of, among the possibilities of the they-self that constantly take it from project to project in the mode of curiosity. But there is also an important second sense in which Da-sein—in the analysis so far—is scattered. Heidegger defined Da-sein's being as care, and care was formulated as being-ahead-of-itself already-in-a-world as together-with-entities. (192) But this means that a structural account of the self of the kind we find in attributionist theories—transformed into a Heideggerian framework—cannot grasp *all* of Da-sein: whatever we grasp, Da-sein is always—ahead-of-itself—still beyond our analysis. In attempting to understand Da-sein as defined by a set of its dispositions or judgments (of by its possibilities), however well these cohere together, we necessarily leave something out, since Da-sein, as *existent*, is not definable by the possibilities it is in at any given time; those possibilities are always

projected ahead. It seems, then, that to understand Da-sein, we will have to grasp its entire life, and this means we have to take up the issue of its death.

There is an obvious problem, however. If Da-sein is always ahead of itself, and we can understand Da-sein only if we grasp it all the way through its death, it seems to follow that we cannot know what Da-sein is until it has died. And this would be problematic if the aim were to give an account of responsibility such that we are responsible for actions that express a *whole* self, since we couldn't know what the whole self is until after death: of course, we might say, an agent's actions *might* express her whole self; but we could not in principle know whether they do until after death.[70] But Heidegger argues that the problem is only a superficial one, based on a misunderstanding of what *death* is. It arises from thinking of death as an event, that will occur at some future point in an agent's existence. And Heidegger responds that once we properly understand death, we will see it as, in a sense, *always* constitutive of Da-sein. This will have an interesting result.

As I've already noted, most accounts of responsibility try to take up some notion of a whole self: the attributionists think of the whole self as made up by the coherence (or incoherence) of an agent's dispositions. Frankfurt thinks of it as constituted by the agent's volitional necessities. Korsgaard takes it to be constituted in the agent's choosing of reasons for actions. And some of these accounts emphasize a diachronic aspect of selfhood: the perseverance of my volitional necessities, the commitment to act for the same reason in relevantly similar future circumstances, the laying out of long-term plans, or the persistence of reasons-responsive mechanisms for which one has already taken

[70] This problem is roughly analogous to Aristotle's difficulty in *Nicomachean Ethics*, I.10, where he wonders how—since happiness depends on one's actions over a lifetime—we could call someone happy while they are still alive.

responsibility, are all attempts to understand the self as a diachronically continuous entity. Heidegger's account, I propose, flips the entire approach on its head. On the one hand, actions do indeed express a *whole* self. On the other hand, this self is not simply diachronically unified: it is always a temporally unified whole. Furthermore, this whole is not constituted by actions, moments of choosing, or any other discrete events. Rather, all events such as deliberate or non-deliberate choices, actions, doings, intendings, belief acquisitions, and so on take place against a backdrop of a temporally unified whole that is their condition of possibility. The whole self—Da-sein's "being-a-whole"—is not dependant on the events that it gives rise to (such as discrete willings or choosing), it is not constituted from within a timeline. Instead, all such acts of the self are constituted from without, by the self's pre-existing temporal unity. And this is what Heidegger's account of death is supposed to give us.

Older readings of Heidegger tend to explain this account of death through some variation on the following theme: knowing that I am going to die, I structure my life according to the recognition of my mortality. Guignon, for example, has repeatedly argued that death, properly understood, imparts a narrative structure on a life. (1984, 2000, 2002) But this cannot be what Heidegger means. First, he insists that death is indefinite; thus, structuring my life in light of it is precisely something I *cannot* do. Second, the reading assumes what Heidegger calls "the vulgar concept of time": at some time later than now, I will no longer exist, and thus am under the imperative of arranging the events of my life—what I did yesterday, what I do now, what I will do tomorrow—in a coherent order that makes sense in light of the expected end. But this is not what Heidegger has in mind. It is the mark of a narrative that it is *going somewhere*; its story

aims at a resolution. But death is not a resolution, or a self-fulfillment or actualization of Da-sein; "for the most part, it ends in unfulfillment, or else disintegrated and used up." (244) And, as I have already hinted, Heidegger simply does not think our life *is* made up of events that can be *arranged* to make up a whole. The point of the narrative interpretation of Heidegger, of course, is that the parts of are constituted by, or make sense in light of, the whole. And this is right. But the reference to "narrative" is then only a misleading metaphor, especially given the role narrative has acquired in recent work as constituting a diachronic self. It is, in other words, precisely the opposite of Heidegger's aim.

More recent interpreters usually spend a good deal of time attacking the older misunderstandings, but when it comes to explaining just what death is supposed to be about, once again talks tends to turn to the way I am to structure my life; perhaps what death reveals is just the greater *gravity* of choosing among my possibilities. Now, it *does* reveal that, in some sense, but *this* doesn't explain what death could have to do with owning oneself, which is clearly to the point, since Heidegger begins Division II, the part of *Being and Time* aimed at making sense of authentic Da-sein, with a discussion of it. For example, Blattner's argument that death be seen as a limit-condition, a boundary on our possibilities, is right. (Blattner 1994) But this interpretation needs to be spelled out in its implications. Death *is* a boundary condition, but what is the implication of this? How does it enable self-ownership? And what does it have to do with our everyday agency if its purpose is not to organize it into a narrative? And, in any case, clearly death is not a boundary condition on just *any* possibility, which is what Blattner suggests when he notes that Da-sein can have an "existential death" at any time—if, for example, I fall out of

love, I can no longer press into my self-projected possibility as a lover. But surely death is not supposed to mark the boundary of each possibility individually, but of *all* possibilities taken together; it is the limit-situation of the possibility of *being-possible* as such: that is, it limits Da-sein's existence as a whole.

Let me note two points about the structural role of death in *Being and Time*, both of which I have already hinted at and will now go on to sketch in more detail. First, it is introduced explicitly in answering the question of how we can grasp the being-a-whole of Da-sein. And second, it makes its return in the text (despite minor mentions) only once Heidegger has introduced resoluteness and notes that, to be fully authentic, it must also be *anticipatory*. In re-introducing this notion, Heidegger notes, furthermore, that discussion of anticipation had, previously (that is, before the introduction of resoluteness), been entirely formal, or a mere "ontological project." (309) What is it that is new in this mention of anticipation in relation to the previous discussion? Clearly—especially given the fact that the section is in prelude to Heidegger's introduction of temporality—what is key about death is that, as a kind of break-down of Da-sein's being-possible, it makes clear the phenomenon of temporality that underlies Da-sein's care structure. And it makes the phenomenon clear precisely through bringing death and resoluteness together as Da-sein's being-as-a-whole projecting itself onto guilt. Thus, we should not focus too much on death as such, but on what it discloses: the being-a-whole of Da-sein, and the notion of future as *anticipation*, which is implied by and first makes sense of guilt. How is this supposed to work?

First, Heidegger takes up the everyday understanding of death as the *end* of Da-sein, in the sense of concluding its life. This understanding turns out to be flawed in a

number of ways. First, it presents death as a common, or public phenomenon. Other people die, and we ourselves will die as well. But this understanding of death is entirely third-personal, and it misses the asymmetry between the death of others and one's own death. We can see others die, but we cannot experience ourselves die in the same way. Thus, death is an individualizing notion: only *I* can die my death, and "no one can take the other's dying away from him." (240) This is what allows death to separate Da-sein from its they self. Among the possibilities laid out by the they, everyone is interchangeable or representable, as already noted. I can grade my students' papers, but another can also do it, and this is due to the public nature of the they's possibilities. But no one can represent my death, and thus no one can represent the being-a-whole that death signifies; "in 'ending' and in the totality thus constituted of Da-sein, there is essentially no representation." (240) Death thus discloses the mineness of existence: my death is essentially not shareable. And since death is supposed to have something to do with being-a-whole, it can therefore disclose Da-sein's totality as not shareable, as a possibility that is not laid out by the they self.[71] Heidegger thus calls death Da-sein's *ownmost* possibility. The point of this slightly misleading formulation is not that dying is the only thing we can do apart from the they self, but that self-ownership—wresting one's authentic self away from the they self—is possible only *in light of* death.

Death is furthermore nonrelational. In taking care of things, Da-sein uses them in order to accomplish something, and what it accomplishes is for the sake of beings like itself. This referential framework constitutes the world. But death has no such relational

[71] Of course the possibilities involved in dealing with death, or preparing for it—making funeral arrangements, meeting with friends and family, writing out a will—are laid out by the they. But death itself is not.

structure: it is not and cannot be *for* anything.[72] One might, of course, organize one's life in a certain way in light of death; but this does not involve using death in-order-to organize one's life in that way; rather, it involves organizing one's life for the sake of death. Da-sein's possibilities typically refer to something else, to an in-order-to or a for-the-sake-of that allows Da-sein to press on into those possibilities. But death is not the sort of possibility that allows pressing on into it. It is both constituted nonrelationally, and is also the possibility one cannot go beyond; it "reveals itself as the *ownmost nonrelational possibility not to be bypassed*." (250-251) But this talk of death as a possibility is especially strange, since possibilities are characterized by our pressing into them; death, on the other hand, is "the possibility of the absolute impossibility of Da-sein," or "the possibility of no-longer-being-able-to-be-there." (250) We understand both ourselves and the world in terms of the possibilities we project. Thus, understanding death as a possibility requires understanding ourselves in light of it, that is, in light of our ownmost, nonrelational, unsurpassable possibility.

Heidegger stresses what should already be clear: that death is, here, an existential concept rather than a biological one. He explicitly invents a terminological distinction to make this point: animals perish, humans demise, but only Da-sein—as the *being* of human beings—dies. So just as Heidegger is at pains to distinguish an authentic understanding of death from inauthentic understanding, he wants to distinguish death as an existential concept from any feature that could belong to things present-at-hand. This is behind Heidegger's insistence that the initial problem—that of understanding Da-sein as being-a-whole before it has ended—is mistaken; death is not "something outstanding,"

[72] This again shows why death is individualizing: since Da-sein normally understands itself in terms of the things it takes care of, in understanding itself in terms of death it is stripped of those ways of self-understanding.

something that is yet to happen. Heidegger contrasts death with ways in which present-at-hand things can end: rivers might reach their completion in the sea or come to a man-made block, weather patterns may end and so may debts, when one discharges them. Once one reaches the end of the river, one has traversed the entirety of that river; and once one pays off a debt by giving what is owed, it is no more. Death can also be seen this way: we *know* that we will die, that our process of life will at some point end. But this kind of empirical certainty, Heidegger suggests, is the way the they tries to cope with death by covering it over: it portrays death as certain, but only in the sense that it "happens," it is an *event* that one undergoes, and that will at some point in the future happen to us all.

In contrast to this empirical certainty of death as the end of a present-at-hand entity, Heidegger suggests the analogy of a ripening fruit. An unripe fruit has not become ripe, but not simply in such a way that the ripeness is an event in its future. The ripeness is not something foreign to the fruit, but is its own completion, and so the unripe fruit does not simply have a state different from that of the ripe fruit—we miss something, for example, if we think of the green tomato and the red tomato as two unrelated event-states of tomatoes. In its ripening, the fruit "is not only not indifferent to its unripeness as an other to itself, but, ripening, it *is* the unripeness. The not-yet is already included in its own being, by no means as an arbitrary determination, but as a constituent. Correspondingly, Da-sein, too, is *always already its not-yet* as long as it is." (244) The unripened fruit, in other words, exists *as* unripe, so that its future finished state of ripeness is already constitutive of what it is as an unripe fruit. Of course Heidegger immediately distinguishes Da-sein from the fruit, since the fruit, unlike Da-sein, reaches

its *fulfillment* in its ripening. Da-sein's death, as noted above, is not a fulfillment in any sense. Nevertheless, this is perhaps the clearest analogy Heidegger gives to explain how he sees Da-sein's relation to its death: as existing, Da-sein is always constituted by its being-toward-death.

Now we can work out the idea that death is a *possibility*, though a rather odd possibility of the impossibility of being. When death is not being covered over as empirically certain, Da-sein can *be* in the certainty—that is, instead of simply knowing that it will die, it understands itself as finite. On the one hand, as we've just seen, this means that death becomes *constitutive* of Da-sein's being-possible; on the other hand, it becomes clear that death—not as an outstanding end, but as thus constitutive—is not an event, but itself a possibility in terms of which Da-sein understands itself and exists. Death is a possibility not in the sense that when its heart stops beating, Da-sein will still be projecting its understanding one last time, but in the sense that Da-sein's possibilities are—as such—limited by and thus understood authentically in light of its ownmost possibility. In other words, death, as being-toward-death, is a meta-understanding: it is Da-sein's understanding of its possibilities, and this is itself its ownmost possibility.

So how does this possibility work? Again, recall that Da-sein's other possibilities are relational, involving an in-order-to and a for-the-sake-of-which. Consequently they are subject to the danger of being interpreted by the they as *mere* possibilities, that is, as processes on the way to actualization. The possibility of being a job-seeker, for example, reaches its actualization in finding a job, and the possibility of being a dissertation writer finds its actualization in a dissertation. "Being out for something possible and taking care of it has the tendency of *annihilating* the *possibility* of the possible by making it

available" (261), Heidegger warns. And this makes it seem as if, in completing a task or carrying out an action, we cease to project possibilities and return to what is actual, as if actuality is the stable state and possibility an occasional distraction from it. This tendency to make the possible real is, after all, a mark of Da-sein's felling prey in curiosity. But what makes such understanding inauthentic is that it overlooks the fact that all actions and tasks are still meaningful only within a further possibility. "The actualization of useful things at hand in taking care of them (producing them, getting them ready, readjusting them, etc.), is, however, always merely relative, in that what has been actualized still has the character of being relevant. Even when actualized, as something actual it remains possible for…, it is characterized by an in-order-to." (261)

One can actualize a possibility only against the backdrop of a further possibility, and an authentic understanding thus sees possibility as higher than actuality—as both prior to it and as a condition of its possibility, not merely in the sense that every actuality is the *outcome* of pressing into some possibility, but in the further sense that every possibility—whether or not it gives rise to actuality—is itself constituted by possibility in light of which the actuality can be what it is. And this is precisely what death gives us to understand, because, as the possibility of the impossibility of existence, it "gives Da-sein nothing to 'be actualized' and nothing which it itself could *be* as something real." (262) In being-toward-death, Da-sein understands itself as essentially being-possible. While other possibilities have some purpose, some end *for which* they are possible, death provides no such purpose. Being-toward-death opens the way for self-ownership because in it Da-sein cannot interpret itself in terms of world and the purposes of the they; it no longer remains fragmented, but is unified *as* possibility in relation to its limiting

nonrelational possibility. In recognizing itself as possibility, Da-sein is freed of its tendency toward entanglement and "from one's lostness in chance possibilities urging themselves upon us, so that the factical possibilities lying before the possibility not-to-be-bypassed can first be authentically understood and chosen." (264) In revealing Da-sein *as* possibility, death frees Da-sein with regard to its factical possibilities: Da-sein sees these not as necessities, but as matters of choice.

This understanding of death involves its own kind of temporality. In everydayness, Da-sein primarily relates to its future through expectation: it treats the future as consisting of events, which will at some point be actualized. By presenting death as such an event, something to merely expect or wait for, the they covers over the authentic understanding of it. As we have seen, the authentic understanding involves recognizing oneself existing as possibility—as being ahead of itself, Da-sein does not have a closed future that it must expect, but an open future that, in principle, cannot be actualized. Heidegger terms this relating to possibility *anticipation*.[73] In expecting, Da-sein sees its future as fixed, consisting of an actualization as a corpse. In anticipation, Da-sein recognizes its future not as an event, but as a mode of being in which its possibility is not actualizable. Anticipating, Da-sein can avoid being entangled in the world and letting its possibilities be dictated to it by its pre-existing commitments to and self-

[73] "Anticipation" is Stambaugh's translation of "Vorlaufen," literally "fore-running," which Macquarrie and Robinson had translated awkwardly as "running ahead in thought." In some ways, anticipation is an unfortunate translation—it obliterates the *active* dimension of *Vorlaufen*, which is obviously crucial to Heidegger's account. On the other hand, *Vorlaufen* has connotations of "preparedness," which suggests a standard Heideggerian theme: the unify of activity and passivity, or perhaps the casting of the active *as* passive; in any case, it is clear that he thinks the common way of drawing the distinction is misleading. Anticipation *does*, however, serve as a nice contrast with "waiting" and "expecting." I will continue using Stambaugh's translation while noting that the active dimension of anticipation needs to be kept in mind. The translation is not a good one; but there are no good English translations of many of Heidegger's key terms. Bringing Heidegger into a serious dialogue with Anglophone philosophy will require a wholescale retranslation and appropriation of his terminology, much as his own appropriation of Aristotle plays a major role in his work. But *that* is obviously not a project I can undertake here.

understanding in the they; "Da-sein guards itself against falling back behind itself, or behind the potentiality-for-being that it has understood." (264) That is, Da-sein *can* no longer understand itself in terms of its prior understanding, because its understanding, as projection of possibility, does not have any particular actualization as its end; rather, Da-sein must constantly exist as possibility, that is, understanding anew and transforming its prior understanding.[74]

Finally, we can bring the results together. Since death is not something outstanding that Da-sein must wait for, we can resolve the initial problem—that the whole self cannot be grasped during life. Death is a possibility of understanding oneself, and as anticipation it already modifies Da-sein's pressing forward into possibilities. "The movement toward a future ability to be constitutes our current ability to be." (Nicholson 2005 55) Da-sein exists as a temporally unified whole not by virtue of diachronic agency—which can, in any case, allow for only a partial and contingent unity—but as anticipating, and therefore understanding itself as possibility. On the one hand, it does not constitute itself as unified by making choices; rather, it makes choices on the basis of a pre-existing unity. On the other hand, it does not always reside in the same volitional necessities, but—understanding them as possibilities—is capable of "betraying" them when called to do so by the situation. And in so doing Da-sein remains the same self, because its unity is not contingent on those necessities, but rather first allows them to be possibilities that Da-sein presses into. Finally, Da-sein exists as a whole in anticipation because it is no longer scattered among possibilities. "Because anticipation of the possibility not-to-be-bypassed also discloses all the possibilities lying before it, this

[74] Velleman's account—on which (full blooded) action is guided by a self-understanding—would on this scheme ensure that *no* action is ever possible: understanding is never finished; in acting on my understanding, I transform that understanding.

anticipation includes the possibility of taking the *whole* of Da-sein in advance in an existentiell way, that is, the possibility of existing as a *whole potentiality-of-being*." (264) Since anticipation discloses *all* possibilities *as* possibilities—because all of them are limited by death, and thus none have any essential privilege or salience in the face of the ownmost possibility—it can bring them together under a single limiting condition that individualizes and unifies Da-sein. In anticipation, then, Da-sein is existentially unified and set free to choose among its factical possibilities, so that its ownmost possibility—and not those factical possibilities—guides its actions. Its agency can thus become its own.

E. Anticipation and Temporality

We must now bring the threads together. Heidegger argues that resoluteness—as authentic being a self—becomes fully authentic only in anticipation. Thus, authenticity requires *anticipatory resoluteness*, which he insists does not involve two phenomena haphazardly brought together, but rather a "modalization" of the latter by the former. In some ways it is already clear how the two belong together: in *Angst*, Da-sein discloses itself in its being-toward-death, since *Angst* presents the world and its possibilities as lacking in salience. And from *Angst*—that is, from anticipation—Da-sein calls to itself in its they-self and into a projection of its being-guilty. But to make sense of how this works, how anticipation *makes sense* of guilt (or, rather, its disclosure of Da-sein as having always already chosen), and how the two together allow for genuine ownership, I

will first draw on Heidegger's account of temporality.[75] I will not work out the account in its full details. Instead, I will focus on the salient points needed to establish my argument: that freedom and responsibility require what I have been calling deep temporality.

It is important to note that, despite frequent emphasis on Da-sein's finitude in the secondary literature, in his discussion of death Heidegger does not do what one would expect: he does not use death to emphasize the distinction between the finite and the infinite. In part, this is because something like infinite understanding plays no role in Heidegger: bringing it in could only serve to cover up the question of being, since the meaning of being is to be disclosed on the ground of temporality. If finitude plays any major role in Heidegger in *contrast* to infinity, it is not in the account of death, but in that of conscience, where guilt does serve the role of characterizing Da-sein as essentially limited in its possibilities. But even there, as I've noted, he immediately stresses that guilt is a *positive* rather than a negative characterization of Da-sein; it is what *allows* Da-sein to *be* something at all.[76] In his account of death, Heidegger is concerned primarily—

[75] Heidegger, of course, works out the unity of anticipatory resoluteness first, in order to disclose the basis for authentic temporality, which allows him to repeat the account given up to that point and leading into his conception of historicity. My purpose here is more modest, and I can thus reverse the order of presentation so as to show how the account of authentic temporality allows for and makes sense of the idea that we have always already chosen our possibilities.

[76] I am therefore puzzled by analyses that claim that, in guilt, Da-sein always falls short of a standard of achieving itself. See, for example, Dreyfus and Rubin's early account of what *Angst* discloses in Dreyfus (1991), and a very different and fascinating account by Tanzer (2001). In the context of defending the claim that Heidegger's account is not purely decisionist and thus inviolable, but rather postulates a norm that *can* be violated, Tanzer emphasizes Da-sein's inability to authentically achieve itself, arguing that Da-sein's guilt involves a constant violation of a norm. On my reading these accounts are contrary to Heidegger's intention: he explicitly states that he is taking over the notion of "guilt" from the common view of what consciousness discloses, and immediately goes on to characterize it as a positive phenomenon, thus in direct opposition to the common view. True, Da-sein cannot achieve itself in the sense of becoming actual; it can only—in death—become not-possible, and this is no self-achievement. But on my reading, Heidegger is not positing a norm that Da-sein always fails to fulfill; rather, he is insisting that the very idea of self-achievement involves an inauthentic understanding of Da-sein, and he contrasts it with the positive understanding of Da-sein as possibility which cannot be actualized. Its aim is not to actualize itself by overcoming possibility, but to actualize itself—if the term still makes sense—by understanding itself *as* being-possible.

almost exclusively—not to distinguish finitude from infinity, but rather to distinguish the existential concept of death from its constant covering over in the they. When he does emphasize finitude, Heidegger does so precisely to separate death from demise: Da-sein "does not have an end where it just stops, but it *exists finitely*." (329) His aim in understanding death existentially is to bring out the notion of anticipation in opposition to expectation and waiting.

What is wrong with waiting for death? One can wait only for an event, but death properly understood is not an *event* at all—it is Da-sein's openness to its existing as possibility. The notion of death as an event belongs to what Heidegger calls the vulgar understanding of time: "What is characteristic of the 'time' accessible to the vulgar understanding consists, among other things, precisely in the fact that it is a pure succession of nows, without beginning and without end, in which the ecstatic character of primordial temporality is leveled down." (329) This, of course, is what I have been calling shallow temporality: a view according to which both our choices and our underlying dispositions and attitudes—our wills—are characterized as *events* happening at a particular point—a now—on a timeline. Heidegger does not insist that this view of time is *wrong*; he simply notes that it is derivative of a more primordial temporality. (326) Understanding Da-sein in terms of a succession of nows is to understand it as a present-at-hand entity, and Heidegger's entire analysis is aimed at showing that such an understanding is inappropriate to the sort of being that Da-sein is. In what sense, then, is the view of time as consisting of a series of events, or nows, not primordial? In his evaluation of Kant's Second Analogy, according to which every (phenomenal) event occurs in accordance with the law of cause and effect, Heidegger notes that causality

already presupposes a deeper account of temporality, since "*perceiving an event* means not just perceiving something as it occurs, but knowing in advance that this follows on from something earlier." (Heidegger 2002 124) Thus, perceiving events as essentially determined involves, in advance, understanding them on the basis of a past occurrence, so that "*causality (as causation) means: running ahead in time as determining letting follow on* such that what runs ahead is itself an event that refers back to something earlier that determines it. As such a relation, causality necessarily involves the temporal character as this going before." (2002 131) Causality is thus not a primordial understanding of the relation between events; it is grounded in anticipation that repeats the past.

But this account does not apply only if we assume that all events are caused by prior events; understanding something as an event already implies a deeper temporal structure. As Heidegger has already argued, our intentionality is grounded in our taking care of things. But taking care implies that we already "retain" a referential whole within which the things can be used and "await" or "expect" a purpose to be attained by their use. "If heedful association were simply a succession of 'experiences' occurring 'in time' and if these experiences were 'associated' with each other as intimately as possible, letting a conspicuous, unusable tool be encountered would be ontologically impossible." (355) Awaiting and retaining are conditions of possibility for experiencing an event as an event; they co-constitute our encounter with things as present or as occurring in a now. So time cannot, primordially, be a series of nows strung together, since the occurrence of each now itself requires a prior understanding of a pre-given referential whole and an expected effect. That we are not normally aware of this underlying temporality is not strange: we do not, in acting, need to explicitly or thematically pay attention to the

referential whole or the context in which we act. Absorption in everyday tasks in fact presupposes a "forgetting" of the whole in order to focus on what we are doing. And, for that matter, "simply looking," or grasping things thematically, involves forgetting the practical context that such looking presupposes. Thus Heidegger has both an argument against the primordiality of shallow temporality, and an error theory for explaining why we normally think that time is primarily a series of nows.

What we have so far—a temporality characterized by "retaining" and "awaiting"—is all that the pragmatist reading of Heidegger typically gives us. But these terms still belong to the vulgar understanding of time: they are appropriate only for understanding our experience of innerworldly things. Authentic temporality—that is, the temporality in which Da-sein fully understands and owns *itself* rather than losing itself in the world—requires finding the grounds of such retaining and awaiting.[77] Awaiting involves some actuality, some concrete, objective event: one awaits only something that can, at least potentially, occur in a future present. Similarly, one retains (or forgets) something that is already there, that itself is objectively present. And one acts, when one acts irresolutely, on the basis of a fixed framework and for the sake of a purpose that this framework allows or affords. Inauthentic temporality, then, is geared toward our dealings with things. But it already demonstrates the impossibility of a present, a now, as a basic constituent of our experience of entities and events. And it already displays a unified structure: "the making present that awaits and retains constitutes the familiarity in accordance with which Da-sein 'knows its way around' as being-with-one-another in the

[77] That authentic temporality, the temporality in which Da-sein understands itself as itself, must ground the inauthentic temporality in terms of which it understands its world and itself as worldly is not surprising. Our relation to the world is permeated by possibility, which is fully grasped *as* possibility only in authenticity. Since our understanding of things in everyday use depends on our projection of possibilities, it follows that everyday temporality will involve a modification of an underlying authentic temporality.

public surrounding world." (354) What remains is to bring out *authentic* temporality and clarify anticipatory resoluteness on its basis.

In his discussion of death, Heidegger contrasts waiting—an inauthentic understanding of death—with anticipation, which is the authentic understanding. Anticipation does not involve expecting an event to happen, but "the *being toward* one's ownmost, eminent potentiality-of-being." (325) That is, the authentic future—the future in terms of which Da-sein understands itself in a way appropriate to an *existing* being rather than a being that is present-at-hand—involves understanding oneself as pure, open possibility that does not aim at actualization. As being-possible, Da-sein is always ahead of itself, but "the 'ahead' does not mean the 'before' in the sense of a 'not-yet-now, but later.'" (327) That something might occur at a later time is obviously not excluded by this notion of the future; the point is only that the sense of "future" in which events happen—the vulgar understanding—is both inappropriate to understanding beings like Da-sein, and requires a sense of future as anticipation as its condition of possibility. The authentic past, similarly, is not simply something one retains or forgets, and is not a "no-longer-now, but earlier." (327) Instead, it is a "having been," which involves coming back to oneself, "back to thrownness as something to be *possibly retrieved*." (343) And of course there is also a "making present," which is unified with the anticipation and retrieval of future and past. Authentic making present, which goes along with anticipation and retrieve, is the Moment (*Augenblick*). All these "ecstases," as Heidegger calls them, are unified in one structure; they are ecstatic, or "stand out from themselves," insofar as each implies the others: the authentic future has a past and a present. The same goes for the other ecstases:

> Understanding is grounded primarily in the future (anticipation or awaiting). Attunement temporalizes itself primarily in having-been (retrieve or forgottenness). Falling prey is temporally rooted primarily in the present (making present or the Moment). Still understanding is always a present that 'has-been.' Still, attunement temporalizes itself as a future that 'makes present.' Still, the present 'arises' from or is held by a future that has-been. (350)

The unity of temporality, in which past, present, and future always belong together, is for Heidegger the meaning of—what makes it possible to project—the care structure. Being-ahead-of-itself as already-being-in-a-world together-with-entities is unified by the unity of temporality, the future of being-ahead-of-itself, the past of already-being-in-a-world, and the present of together-with-entities. And this temporal unity can be either inauthentic, geared towards grasping innerworldly beings as present in a now, or authentic, geared toward Da-sein's self-understanding as being-possible. This unity is supposed to make sense of the structure of anticipatory resoluteness. Heidegger insists that resoluteness must project its being-guilty onto Da-sein as a whole; and Da-sein as a whole is grasped in the mode of anticipation. But this is not particularly clear. Let me take up two further hints about the relation between anticipation and resoluteness. On the one hand, anticipatory resoluteness is "the understanding that follows the call of conscience and that frees for death the possibility of *gaining power over* the *existence* of Da-sein and of basically dispersing every fugitive self-covering-over." (310) Since anticipation involves understanding oneself as possibility, this self-understanding "gains power" over Da-sein in the sense of letting Da-sein own itself, or taking over its ownership from the they-self. On the other hand, "being guilty, which is constantly with us, does not show itself without being covered over in its character as prior until that priority is placed in the possibility which is for Da-sein absolutely *not to be bypassed*." (307) In other words, being guilty—which I have characterized as disclosing that we have

always already chosen—can only make sense as such, without being misinterpreted, in light of anticipation of death.

So now we come to the crux of the issue: how does anticipation "modalize" resoluteness in such a way as to disclose ourselves as having always already chosen? In the previous discussion of this idea, it seemed that this notion of choosing is only a metaphor; that, after all, we have *not* chosen, but the they has chosen for us. And it seemed like we cannot make sense of the idea that we have *let* the they "disburden" us of our choice of possibilities; there seems to be no room for agency in this picture. This objection assumes that the past is absolutely fixed, as something that was and is now no longer. That is, it assumes inauthentic temporality. But Da-sein's *authentic* temporality includes the past *in* the future, and the future in the past. We project ourselves always out of thrownness, but we also retrieve our thrownness in the project. "The authentic coming-toward-itself of anticipatory resoluteness is at the same time a coming back to the ownmost self thrown into its individuation. This ecstasy makes it possible for Da-sein to be able to take over resolutely the being that it already is." (339) So in coming forth to itself as its ownmost possibility, Da-sein *also* comes back for itself in its thrownness. Resoluteness projected Da-sein onto its guilt in thrownness, and placed Da-sein under an imperative of taking over its ground and existing out of it. So how does Da-sein do that?

Anticipation discloses Da-sein as a whole. This means not only that Da-sein is constituted by its future so that its future is not something still outstanding, but also that its past is never something that has simply happened. "To take over thrownness means to authentically *be* Da-sein in the *way that it always already was*." (326) As being-a-whole, Da-sein is not simply its present, with a past that has already happened and is, in a way,

already closed off as a possibility. As existing, Da-sein always exists *as* its possibility—that is, it *is* what it understands itself as in acting. Da-sein's facticity, then, is not something simply given that it must accept, but something offered that it must take up as also constitutive of its possibility in its projection of itself onto its ownmost possibility. The imperative to take over one's ground, to understand it within the context of one's ownmost possibility, comes from that ownmost possibility itself: existing as *possible* and never actual, Da-sein is not merely opened up to its possibilities in the present, but must take up its past as part and parcel of its pressing into its open possibility. Consequently, "Da-sein can *be* authentically having-been only because it is futural. In a way, having-been arises from the future." (326) Anticipatory resoluteness sets Da-sein free from determination by the they and by its past precisely because, in Da-sein's being-a-whole, its future is constitutive of its past. In disclosing itself as possibility, separated from the distortions introduced by the they in which Da-sein tends to understand itself as a mere thing, Da-sein can understand its past as its own, that is, as chosen on the basis of its open future. Of course this does not change "the facts" of Da-sein's past; but "facts" enter into Da-sein's constitution only as facticity, that is, as involved in but not determining of its self-understanding. In understanding itself, Da-sein understands itself as free or, as Heidegger puts it, "understanding the call, *Da-sein listens to its ownmost possibility of existence*. It has chosen itself." (287)

But isn't Da-sein still bound by the they? There is a lively debate on this topic, as there is on almost every aspect of Heidegger. Dreyfus (1991) used to insist that Da-sein is always trapped in the they.[78] Heidegger does, after all, tell us that "*authentic being one's self* is not based on an exceptional state of the subject, a state detached from the they, *but*

[78] As I will mention below, Dreyfus's view on the topic has changed.

is an existentiell modification of the they as an essential existential." (130) And this seems to suggest that Da-sein is *always* trapped in the they, so that authenticity does not free it after all. To some extent, this view has been helpfully fixed by Guignon (1984) and, in detail, by Boedeker (2001), who note that authenticity involves a grasp or modification of the *they*, but it is opposed to the *they-self*. The idea, then, is that we can free ourselves from the they-self in the individuation of *Angst*, but the they remains the source of our possibilities. Boedeker thus argues that "Dasein in the self-owning mode of Being-itself thus projects the same concrete possibilities of itself as it does in the mode of the *Man*-self. What is distinctive about the mode of self-ownership is that Dasein for the first time owns up to the existential consequence of doing so imposed by its ownmost possibility of death." (2001 89) But Heidegger also speaks of authentic possibilities, and of Da-sein being led astray from those in the they (174, 178, 344, see also Bracken (2005)). This seems contradictory: either Da-sein's possibilities are entirely drawn from the they, or Da-sein introduces possibilities of its own.

The answer, I think, may be that both are right. Boedeker suggests that in choosing in light of death, Da-sein *is* taking up a different possibility: after all, understanding itself authentically is *itself* a possibility, and perhaps this is what Heidegger means; but then it becomes unclear why he might speak of authentic possibilities in the plural. Perhaps what Heidegger means is, rather, something like the following. Return to the example of walking on a sidewalk. I walk on the sidewalk *because* that is what one does. But I can also walk on the sidewalk—doing what one does—because I choose to. Perhaps, then, Da-sein is open to authentic possibilities in the sense that it can take up the same old possibilities of the they, but understand them—and

so exist—in a new way, exercising a new possibility. This may be what Boedeker has in mind. But this also doesn't seem quite right, because it clashes with Heidegger's account of falling prey. Heidegger notes, for example, that curiosity—constantly flitting here and there—never dwells anywhere; that idle talk presents everything as understood and thus does not have time for authentic understanding. Fallen Da-sein does not seem to *act* in quite the same way as authentic Da-sein.

In an pessimistic sounding moment, Heidegger notes that, "Da-sein can never escape the everyday way of being interpreted into which Da-sein has grown initially. All genuine understanding, interpreting and communication, rediscovery and new appropriation come about in it and out of it and against it. It is not the case that a Da-sein, untouched and unseduced by this way of interpreting, was every confronted by the free land of a 'world.'" (169) But the public way of interpreting does not *allow* for anything genuinely new; thus, Heidegger is suggesting that something new *is* possible. But it is immediately taken back up into the they, so its newness is quickly covered over by idle talk; and it appears only against an existing shared backdrop of the they. But this is not strange: Da-sein's world is a public world. To entirely escape the they, Da-sein would have to escape the world; and then it would no longer be Da-sein. But to say that every interpretation and understanding is grounded in the they and returns to it is not to say that nothing new is possible. Da-sein *can* authentically take up possibilities that draw on, but are not entirely drawn from, the possibilities of the they. This is precisely the point: the they, as having-been, can provide a *ground*—and, indeed, an imperative—for taking up authentic possibilities. One's facticity may always be characterized by the they; but insofar as we can choose new possibilities on the ground of that facticity, we also choose

the facticity as the ground of those possibilities. There can be no existing into the future without a past, which is at least part of what conscience discloses. Only in inauthenticity can Da-sein flit from possibility to possibility detached from a past it forgets; but it is only *when* it does so that it fails to *retrieve* its having-been and it thus lets itself be determined by that past.

F. The Situation and Self-Ownership

Finally, there is an interesting puzzle about what anticipatory resoluteness calls us to. Heidegger insists that resoluteness brings us to action, but it might seem that the action in question is meaninglessly abstract. In response, Heidegger insists that "the call of conscience does not dangle an empty ideal of existence before us when it summons us to our potentiality-of-being, but *calls forth to the situation*," which resoluteness both discloses and places itself into. (300) What he says concerning the situation is not entirely clear. We learn, for example, that in the situation Da-sein "becomes free of the entertaining 'incidentals' that busy curiosity provides for itself, primarily in terms of the events of the world." (310) And the situation cannot be available to the they, which "knows only the 'general situation,' loses itself in the nearest 'opportunities,' and settles its Da-sein by calculating the 'accidents' which it fails to recognize, deems its own achievement and passes off as such." (300) What are these "accidents" that the they deems (mistakenly) its own achievement? The suggestion, I gather, is that the they does not know what it is supposed to do in a situation because it approaches each situation entirely on the basis of "calculating," and doing what one does, rather than what is called

for. But there is a more interesting suggestion here, which I will return to shortly: that the they takes its responses to the situation *as its own achievements*, as something it has accomplished, when in fact no one has accomplished anything: the they, seeing only the "general situation," that is, seeing each situation as falling under a type in which *one* does this or that, acts, but the action is unowned.

Heidegger had earlier discussed the situation in relation to *phronesis*, in the context of his account of *Nicomachean Ethics*, Bk. VI. He refers to Aristotle's notion that the *phronimos* acts in the right way, with regard to the right people, using the right things, and so forth. "These circumstances characterize the *situation* in which Dasein at any time finds itself…In this way, Dasein, as acting in each case now, is determined by its situation in the largest sense. The situation is in every case different. The circumstances, the givens, the times, and the people vary. The meaning of the action itself, i.e., precisely what I want to do, varies as well." (Heidegger 1997 100-101) The situation is, there, disclosed by *phronesis*, which guides Da-sein to its resolution or decision and through action. "In every step of the action, *phronesis* is co-constitutive." (1997 101) In *Being and Time*, however, resolution is seen as disclosing the situation, whereas *Umsicht*, Heidegger's translation of *phronesis* in *Sophist*, hardly makes an appearance in this context. Especially since Kisiel's *The Genesis of Heidegger's Being and Time*, these points have led to a common, and not entirely unwarranted speculation that resoluteness is Heidegger's taking up of Aristotelian *phronesis*.[79] But to what extent and in what way?

[79] For a dissenting voice, see Sadler (1996 150), who argues that Heidegger's fundamental disagreement with Aristotle about the relative standing of *phronesis* with regard to *sophia*, his alterations in translation of key terms, and his general tendency to appropriate other thinkers in ways uniquely his own, makes the thesis of *Nic Ethics*'s direct influence on the content of Heidegger's philosophy (as opposed to its method) highly suspect.

Dreyfus (2000a), for example, typically sees resoluteness as *phronesis* in the sense of skill acquisition: the virtuoso is one who always does the right thing within the situation because he has progressed beyond (even internalized) rule application. But Dreyfus now grants that resoluteness allows Da-sein to recognize the contingency of its thrownness and thus go beyond the they in solving problems; anticipatory resoluteness, on the other hand, lets Da-sein recognize—through death—the contingency of the cultural heritage, thus allowing for something radically new. This takes us rather far beyond skill acquisition, and certainly further than Carman's view of resoluteness, according to which "resolute agents… maintain a subtle feel for the situations they confront and so are able to deal with them intelligently, skillfully, with finesse." (Carman 2006 234) This is not wrong, but it seems grossly incomplete; and, as in Dreyfus's account, Carman seems to take the relation between anticipation and resoluteness as more or less contingent, related only by the fact that "what the two notions have in common… is precisely their emphasis on finitude and particularity," allowing Da-sein to avoid being "assimilated into any generic or impersonal conception of people like me in situations like this." (Carman 2005) These accounts fail to do justice to the intimate connection between anticipation and resoluteness (especially since the former is intended to make sense of the latter), the relation of anticipatory resoluteness to temporality (in Carman), and especially its role as an account of self-ownership (in Dreyfus, though Carman's emphasis on first-personal experience strikes me as misleading). But they also, I believe, overlook an interesting textual detail.

Recall that Heidegger describes the authentic future in terms of Da-sein coming toward itself (325) and the authentic past as its coming back (326). But when Da-sein

comes back to itself on its way toward itself, where does it head? "Resolute being together with what is at hand in the situation, that is, letting *what presences* in the surrounding world be encountered in action, is possible only in a *making* that being *present*." (326) The pattern is a repetition of the earlier account of conscience: Da-sein calls to itself in the they, and calls it to what? To the situation! This should look puzzling to commentators, though often it apparently does not. But clearly something odd is afoot: in an account of authenticity, Da-sein's self-understanding which allows it to *own* itself, Heidegger repeatedly refers to Da-sein in the past and future, but brings up the *situation* in the present. Is the situation, then, Da-sein's authentic present? There are, no doubt, some vestiges of Aristotle: in acting, the *phronimos* exercises his virtue, both expressing and maintaining his character. And Heidegger, too, notes that the situation is not merely the context in which Da-sein can act, but rather one in which it is "already acting." (300) And this should remind us of the *Sophist* claim, above, that Da-sein, in acting, "is determined by the situation." Since Da-sein, as care, is always involved with entities, is always circumspectly dealing with them, its acting and its being are co-constitutive.

So what is the point of the situation? Why would Heidegger spend so much space on conscience and anticipation, simply to end up with *phronesis*? Because he wouldn't. Recall that Da-sein necessarily understands itself and world together. It is inauthentic when it understands itself in terms of world; authentic when it understands itself in terms of itself. But this sounds like the exact opposite of being determined by the situation. So long, that is, as we remain internalists, at least of a sort. Understanding itself inauthentically, Da-sein can never dwell anywhere, it is constantly distracted by incidental "new" possibilities; in the they, it acts in an unowned way by simply retaining

what it has been given on the basis of possibilities laid out for it. As authentic, or owned, Da-sein can be determined by the situation, which is unified rather than broken down into pre-calculated types. It is no longer fragmented among the they-dictated facets of the world. "When one is absorbed in the everyday multiplicity and rapid succession of what is taken care of, the self of the self-forgetful 'I take care of' shows itself as what is constantly and identically simple, but indefinite and empty." The self is "unified" only in a completely empty way; but "the *constancy of the self* means nothing other than anticipatory resoluteness." (322) Da-sein has a constant self—we might say it constitutes itself—by taking ownership of its world, temporally unified as a situation.

Unlike a now, which can last forever—is "now" this minute? this day? this year? it's length is entirely indefinite, as Augustine had already noted—the *Augenblick* has no duration; it allows the past and present to meet, so that Da-sein, in taking over the situation that determines its action, can let itself be so determined on the basis of its freely having chosen itself in light of its ownmost possibility. Since its future is entirely open as possibility, and from it it comes back to itself to take up its having-been as its choice, Da-sein is not determined either by the weight of what it retains nor by any definite aim. In letting itself be determined by the situation, Da-sein paradoxically avoids any kind of determinism, because the situation is its own: "It *gives* itself the actual factical situation and *brings* itself into that situation… It is disclosed only in a free act of resolve that has not been determined beforehand, but is open to the possibility of such determination." (307) Thus, acting on its own self-given situation—self-given because it is disclosed entirely without the they—Da-sein frees itself; and relative to the situation it can (and, authentically, must) always retrieve itself anew in light of its anticipation: it

holds itself in the certainty of what the situation demands, and this, "as a resolute holding oneself free for taking back, is the *authentic resoluteness to retrieve itself*." (308) Somewhat paradoxically, "resoluteness is freedom to *give up* a definite resolution, as may be required in the situation. Thus the steadiness of existence is not interrupted, but precisely confirmed in the Moment." (391)

Da-sein, in other words, constitutes itself in anticipatory resoluteness: it takes on a constancy in which it remains faithful to itself and its self-understanding as possibility. Its self-constitution does not bind it to any concrete commitment in terms of which it must then go on to define itself, but rather allows it to be open to the demands of the situation, so that it is always prepared to take back its attachments when needed. It is individualized and whole, so that its actions express its will, the background through which—on the basis of its possibilities—Da-sein responds in acting to solicitations. And its will is unconditioned, because as thrown, Da-sein must take up its ground and exist from it, and it can understand that ground as chosen on the basis of its ownmost possibility, thus making it its own. As whole, Da-sein can make sense of and transform its past—not, of course, in the sense of changing the facts of that past, but in letting its past as having-been determine its future only in light of that future, as meaningful only relative to that future. What openness to the future dictates in the present resolve is, of course, impossible to determine in advance; but it is not empty, either, since it is determined by the concrete situation in which Da-sein, in acting, can be what it is. As Heidegger famously wrote in an approving defense of Kant, the question of what one must will in adherence to the fact of reason is answerable thus: "Everyone who actually wills knows: *to actually will is to will nothing else but the ought of one's existence*." (2002 196)

In willing, as opposed to merely wishing and awaiting in the mode of the they, Da-sein owns itself. The possibility of owning itself is already present in its *always-mineness*, which, as always concerned with its being and under the necessity of becoming itself in self-understanding, has the self-ownership of anticipatory resoluteness as its constitutive aim. In fulfilling this aim—itself disclosed in resoluteness as guilt—Da-sein can first take responsibility by retrieving its having-been from its always already having been chosen. And although it has let the they self choose itself, it can "make up for not choosing" by reclaiming that choice by letting its past be constituted by the open possibility of anticipation and the demand of the situation. At the same time, its ability to self-owningly be determined by the situation rests on a retrieval of its having-been, from which it presses into its ownmost possibility.

To say that Da-sein has a constitutive aim, which it fulfills or fails to fulfill, is to let it be self-created, in a sense, by being toward its future. But this doesn't seem to be the whole story: after all, Da-sein is, for the most part, not authentic. Heidegger frequently notes that Da-sein is inauthentic, irresolute, and lost "initially and for the most part." While commentators differ on whether this means that authenticity can be maintained, or whether authenticity might not rather play a merely methodological role in the account, so that it is not even meant to be a possibility Da-sein can live in (Staehler 2008), Heidegger does seem clear that Da-sein does not "achieve" itself in authenticity and remain perpetually in the moment until its demise. And in any case, authenticity requires understanding the call of conscience, which in turn takes wanting-to-have-a-conscience; someone who wants to have a conscience may become responsible, but what about the rest of us?

Recall that Da-sein has *let* the they take over its choice, not in the sense that the they forced it to choose its possibility, but rather in that Da-sein itself—as a they-self—chose in the mode of the they. That Da-sein has always already chosen itself is *disclosed* in authenticity and covered over in inauthenticity. And only authenticity allows Da-sein to understand itself as being-possible and to take over its ground as having-been. But Da-sein always takes over its ground, because it always exists as possibility, though it may understand itself in terms of innerworldly entities and thus lose itself, forgetting that it is never actuality, and letting the they direct it into frivolous and ever-changing possibilities. In authenticity, Da-sein understands itself as responsible and thereby *takes responsibility*, a process that is *temporal* but not historical: after all, taking responsibility in this sense is not a one-time affair that gives rise to responsibility for future actions. Instead, it is the co-constitutive past dimension of all agency in the Moment, in light of the anticipatory being-a-whole with which Da-sein compares itself.

But taking responsibility in this sense is not a prerequisite for *being* responsible. Instead, it involves *disclosing* oneself as already responsible, as guilty and therefore always under the necessity of defining oneself essentially. This disclosive component of authenticity is the essential step that allows always-mineness to be owned, because disclosing oneself as possible just *is* what is involved in self-ownership. If Da-sein were not always concerned with its being and under the necessity of taking up its thrown ground into its possibility, and if it were not thus responsible for the possibility as which it exists, it could never take responsibility, either. So self-owning action determined by the situation is not necessary for being responsible. Rather, being responsible is the condition of possibility for taking responsibility and owning oneself. On the other hand,

self-ownership *discloses* Da-sein as responsible, which is why it cannot be bypassed in an attempt to explain how responsibility is possible. Authenticity may be necessary for *autonomous* action, insofar as owning ourselves allows for a distinction between what is properly ours and what is not and autonomy requires self-government. Inauthentic Da-sein, failing to make the distinction, is always governed by norms that are not its own because it exists as unowned. But responsibility need not hinge on the exercise of autonomy, despite occasional attempts to define the two as co-extensive.[80]

On the one hand, then, Da-sein is always already pressing forth into its possibilities and existing as these possibilities: it constitutes itself in acting in the world. On the other hand, in understanding itself in terms of itself rather than world, Da-sein takes ownership of itself and defines its being in terms appropriate to the sort of entity it is. Because it is concerned about its being and must define itself through projective understanding, and because its projective understanding is *appropriate* to the kind of being it is only in anticipatory resoluteness, Da-sein's authenticity is a "factical ideal," as Heidegger calls it, or a constitutive aim of its pressing forth into possibilities. This aim, however, is the aim of Da-sein's being as such, which is defined by its so pressing forth, that is, in action. But the aim is not, as in the standard constitutivist accounts, an aim constitutive of individual choices, intentional actions, or practical deliberation. It is the aim of Da-sein's existence as being-possible, on the basis of which choice, action, and deliberation can occur. Of course there is a further question—in the constitutivist mode—about whether any *norms* that are still remotely recognizable as ethical norms can be derived from this aim. But I cannot see why deriving norms from "willing the ought of

[80] Kant, of course, did not think responsibility is co-extensive with autonomous action. One acts autonomously only when acting on the moral law (that is, according to duty and from duty), but responsibility hinges only on our being able to conform to the law.

one's existence" should be especially complicated compared with deriving norms from self-constitution or, for that matter, self-understanding.[81]

Da-sein's self-constitution therefore allows for self-creation in a sense. Its will is unconditioned (though not unlimited) insofar as it is grounded in possibility. In existing as a temporally unified whole—or at least in being able to grasp itself as such—Da-sein *can* act in light of a self-constituted will that it has been in the past and takes up in light of the future, and its present actions and choices are in turn constituted by this prior constitution; the will is not, pace-Korsgaard, dependent on individual moments of choosing, which attempt to establish self-constancy from temporal nows rather than from adherence to a temporal unity in the Moment. The account is thus closer to Frankfurt's defense of self-acquiescence: in action, Da-sein *does* acquiesce to the self it has already chosen in anticipation. But this self is not fixed by volitional necessities, which cannot be violated without ceasing to be the self that it is: being free to take back its commitments in response to the demands of the situation *is* what allows for self-constancy: a self bound to particular volitions or commitments even in the face of good reasons to abandon them lacks constancy because its actions are always guided by necessities rather than by its own appropriation of the context of its action.

[81] I note here that Velleman does not exactly think that we *can* derive ethical norms from self-understanding. Rather, he believes that in aiming for self-understanding, we automatically place ourselves under more or less contingent norms. Moral philosophy is a post-facto attempt to grasp those norms.

7

Conclusion

In the opening chapters, I somewhat roughly lumped attitudes together as an agent's *will*. In turning to responsibility, I first presented the will as involving a background composed of dispositions, attitudes, evaluative judgments, and so on. And, finally, in Chapter 5, I described the will more concretely—though still vaguely—as the articulation (again, not necessarily verbal or explicit) of character, that is, the background of responses to solicitations in the world. In that account, the will was constituted by the possibilities projected by attuned understanding—since it is on the basis of these that affordances can solicit us—but these turned out to be entirely subservient to the they. The question therefore arose of how we can take *ownership* of these possibilities and, thus, of the will. I argued that coherence and conscious deliberation are both unsuitable for providing a genuine account of self-ownership and that even Fischer and Ravizza's approach—despite coming closer—falters on the shallow temporality of its historical view. Let me bring the accounts together.

In my discussion of free will, I distinguished between choices conceived as events, and attitudes making up the will in light of which those choices first *appear* as choices and must be resolved. And I defined deep temporality as an account in which at least one of the attitudes is not—or does not have as a starting point—an event on a timeline. I then argued, or at least attempted to make plausible, that both compatibilism and libertarianism run aground in part because they are temporally shallow accounts: they

present both the choices and their underlying attitudes as events occupying a single timeline. And I suggested that deep temporality coupled with ownership could point the way to a solution to a particularly strange problem that follows from Nagel's eliminativism: that agency seems to be excluded either if the agent's choices are not conditioned by his will, or if his will is itself conditioned by prior events, whether or not those events are choices. The schematic solution to this problem goes as follows.

Our choices are conditioned by our will, since they necessarily take place against a background of possibilities in terms of which we understand ourselves and our world. And our will is chosen, but it is not chosen by the sort of choice, in time, that issues from that will. There is no time in which such a will-forming choice occurs. The point is not that at some instant in time we choose ourselves, but that we have always already chosen ourselves. We know that we have always already chosen ourselves because we are always ahead of ourselves—we exist as open possibility and not as fixed entities with determinate properties. This is disclosed through anticipation, or our openness to death. And since anticipation discloses to Da-sein its being as a whole, it can *retrieve* its initial choice *as* choice. What I have been, in other words, is constituted by what I aim to be. So on the one hand, this account satisfies the condition that the agent's choices must arise from the will rather than the other way around, since what we do in the world depends on the will we have. On the other hand, the will is not conditioned by any prior events. Nor is it already pre-given, as something we find ourselves with and must act in light of: we take up our past in light of the anticipation of the future, not the other way around. Therefore, whatever I have already chosen, it does not necessitate what I do, since my past is part and parcel of my temporally extended will as a whole.

Along with attributionism, we can claim that we are directly responsible for what we do and what attitudes we have regardless of whether or not we have made a conscious choice to so act or to have such an attitude. Our subpersonal mechanisms can be representative or expressive of the self because they issue from the self's projected possibilities, and they are just as representative or expressive of those possibilities as our conscious deliberation which, after all, takes place within the same background. In inauthenticity, of course, we can say that both the subpersonal mechanisms and the conscious deliberations are *reflective* rather than expressive of the agent. But this does not eliminate responsibility: in both cases, world and self are co-constitutive. The difference is in whether the self understands itself and its world in its own terms, or whether it understands itself in terms of world. And while this may make a difference to whether or not an agent *takes* or *accepts* responsibility, it does not make a difference to whether the agent *can* take or accept responsibility, that is, to whether or not the agent *is* already responsible. Thus, attributionism is retained, though substituted with an account of authenticity that allows us to *own* our wills and, moreover, provides a vantage point from which we can see that we are directly (absent either conscious deliberation or history) responsible for our actions and attitudes because we have *chosen* the will that they stem from. The Medievals were right: we are free by virtue of having a will. But we are not free because our will is always in its own power as such; we are free because the will is temporally constituted.

Eliminativism poses no challenge to this account. The problem of eliminativism is that the self, seen as a part of the natural world, seems to simply dissolve into that world on the objective view. What we do appears to be necessitated by the world, leaving us as

agents out of the loop. How can we be free and responsible on this picture? I suggested in Chapter 2 that this question, when asked, necessarily appears as self-deception, and Augustine noted long ago that "the only reason that most people are tormented by this question is that they do not ask it piously; they are more eager to excuse than to confess their sins." (1993 73) *That* in itself does not, of course, mean they are wrong: seeking an excuse perhaps increases the chance that one will make an error in one's own favor, but it does not guarantee such an error. On the other hand, the strength of Heidegger's account lies precisely in his de-moralization of authenticity. To see ourselves as entirely a part of the natural world, constituted by events among other events, is an inauthentic understanding—an understanding of ourselves that is drawn in by worldly entities and casts itself as one of them. The idea that our agency might simply be reduced to the agency of the world, taking away our freedom and responsibility, misconstrues our essence. The authentic self, indeed, does let itself be determined by the world in a sense; but it is precisely when it does so that it is most fully self-determined, because it retrieves itself in the *situation* that determines it.

On a final note, we may ask: even if the Heideggerian account might offer a response to eliminativism, can it really handle a threat to freedom like causal determinism? A thoroughgoing naturalist is likely to say no. But the fact that thoroughgoing naturalists are unlikely to accept that causality itself can only be disclosed to an entity that has the temporal structure of anticipation and retention should not rule out the possibility of offering a response to causal determinism; it rules out only the possibility of convincing some determinists. But with some other determinists, there can be a dialogue. One of the strongest defenders of hard determinism, Ted Honderich, has

recently suggested a model he thinks would allow, in some sense, for a reconciliation between determinism and the sense of our lives as individual and our own. Though Honderich rejects the Kantian noumenal self, he suggests that we may draw on a similarly radical idea, one that involves, roughly, a theory of perception on which perceived objects are, literally, constituted by the atoms making up the objects together with our neuronal structure; this view of consciousness allows for a view of personal consciousness as well, which will depend uniquely on our own neural framework rather than the shared one that contributes to constituting the public world. And even if (or, rather, even though) determinism is true, this view of perceptual consciousness "explains your sense of your life as a sense of something for which you are accountable and also something that is individual." (Honderich 2002 151) *This* idea does not make sense; even if the public world depends on each individual's perceptual consciousness, and each individual's perceptual consciousness is unique, that can explain at best how our lives and worlds depend on our unique neural architecture, but it leaves responsibility entirely unexplained. The strength of the Heideggerian account is precisely that the world does not depend on perceptual consciousness; it depends on our projected possibilities, which involve always *pressing forth* into them. We project our possibilities in agency. And it is in this sense that we can be responsible agents even in a determined world.

References

Anscombe, Elizabeth. 2000. *Intention*. Cambridge, MA: Harvard University Press.

Anselm. 1995. *Monologion and Proslogion*. Translated by T. Williams. Indianapolis: Hackett Publishing.

Aristotle. 2009. *The Nicomachean Ethics*. Translated by D. Ross. New York: Oxford University Press.

Arpaly, Nomy. 2003. *Unprincipled virtue : an inquiry into moral agency*. Oxford ; New York: Oxford University Press.

Arpaly, Nomy, and Timothy Schroeder. 1999. Praise, Blame and the Whole Self. *Philosophical Studies* 93 (2):161-188.

Augustine. 1993. *On Free Choice of the Will*. Translated by T. Williams. Indianapolis: Hackett Publishing Company.

Ayer, A. J. 1982. Freedom and Necessity. In *Free Will*, edited by G. Watson. New York: Oxford University Press.

Bernard. 1920. *The Treatise of St. Bernard Concerning Grace and Free Will*. Translated by W. W. Williams. New York: The Macmillan Company.

Blattner, William D. 1992. Existential Temporality in *Being and Time* (Why Heidegger is not a Pragmatist). In *Heidegger: A Critical Reader*, edited by H. L. Dreyfus and H. Hall. Oxford: Basil Blackwell.

———. 1994. The Concept of Death in *Being and Time*. *Man and World* (27):49-70.

———. 1996. Existence and Self-Understanding in *Being and Time*. *Philosophy and Phenomenological Research* (56):97-110.

———. 2006. *Heidegger's Being and Time: A Reader's Guide*. London: Continuum.

Boedeker, Edgar C. 2001. Individual and Community in Early Heidegger: Situating *das Man*, the *Man*-self, and Self-ownership in Dasein's Ontological Structure. *Inquiry* (44):63-100.

Bracken, William F. 2005. Is There a Puzzle About How Authentic Dasein Can Act?: A Critique of Dreyfus and Rubin on *Being and Time*, Division II. *Inquiry* 48 (6):533-552.

Bratman, Michael. 2001. Two Problems About Human Agency. *Proceedings of the Aristotelian Society* 101 (3):309-326.

Carman, Taylor. 2005. Authenticity. In *A Companion to Heidegger*, edited by H. Dreyfus and M. Wrathall. Oxford: Blackwell.

———. 2006. The Concept of Authenticity. In *A Companion to Phenomenology and Existentialism*, edited by H. Dreyfus and M. Wrathall. Oxford: Blackwell.

Chisholm, Roderick. 1966. Freedom and Action. In *Freedom and Determinism*, edited by K. Lehrer. New York: Random House.

———. 1982. Human Freedom and the Self. In *Free Will*, edited by G. Watson. New York: Oxford University Press.

Clarke, Randolph. 1993. Toward a Credible Agent-Causal Account of Free Will. *Nous* 27 (2):191-203.

———. 2005. On an Argument for the Impossibility of Moral Responsibility. *Midwest Studies in Philosophy* XXIX:13-24.

Crowell, Steven. 2007a. Conscience and Reason. In *Transcendental Heidegger*, edited by S. Crowell and J. Malpas. Stanford: Stanford University Press.

———. 2007b. *Sorge* or *Selbstbewußtsein*? Heidegger and Korsgaard on the Sources of Normativity. *European Journal of Philosophy* 15 (3):315-333.

———. 2008. Measure-taking: Meaning and Normativity in Heidegger's Philosophy. *Continental Philosophy Review* (41):261-276.

Davidson, Donald. 1980a. Actions, Reasons, and Causes. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980b. Agency. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980c. Freedom to Act. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980d. How is Weakness of the Will Possible? In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980e. Intending. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980f. Mental Events. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 1980g. Psychology as Philosophy. In *Essays on Actions and Events*. New York: Oxford University Press.

———. 2004. Problems in the Explanation of Action. In *Problems of Rationality*. New York: Oxford University Press.

Dreyfus, Hubert. 1991. *Being-in-the-World*. Cambridge, MA: MIT Press.

———. 2000a. Could Anything be More Intelligible than Everyday Intelligibility? Reinterpreting Division I of *Being and Time* in the Light of Division II. In *Appropriating Heidegger*, edited by J. E. Faulconer and M. Wrathall. Cambridge: Cambridge University Press.

———. 2000b. Responses. In *Heidegger, Authenticity, and Modernity*, edited by M. Wrathall and J. Malpas. Cambridge, MA: MIT Press.

———. 2005. Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise. *Proceedings and Addresses of the American Philosophical Association* 79 (2):47-65.

———. 2007a. Detachment, Involvement, and Rationality: Are We Essentially Rational Animals? *Human Affairs* (17):101-109.

———. 2007b. Response to McDowell. *Inquiry* 50 (4):371-377.

———. 2007c. The Return of the Myth of the Mental. *Inquiry* 50 (4):352-365.

Ekstrom, Laura Waddell. 2000. *Free will : a philosophical study*. Boulder, Colo.: Westview Press.

———. 2005. Alienation, Autonomy, and the Self. *Midwest Studies in Philosophy* XXIX:45-67.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control*. New York: Cambridge University Press.

Frankfurt, Harry. 1969. Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy* 66 (23):829-839.

———. 1982. Freedom of the Will and the Concept of A Person. In *Free Will*, edited by G. Watson. New York: Oxford University Press.

———. 1992. The Faintest Passion. *Proceedings and Addresses of the American Philosophical Association* 66 (3):5-16.

———. 1999a. Autonomy, Necessity, and Love. In *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

———. 1999b. On the Necessity of Ideals. In *Necessity, Volition, and Love*. Cambridge: Cambridge University Press.

———. 2002. Reply to J. David Velleman. In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by S. Buss and L. Overton. Cambridge, MA: MIT Press.

———. 2006. *Taking Ourselves Seriously & Getting it Right*. Stanford: Stanford University Press.

Freud, Sigmund. 1960. *The Ego and the Id*. Translated by J. Strachey. New York: W. W. Norton & Company.

Ginet, Carl. 1990. *On Action*. New York: Cambridge University Press.

———. 2007. An Action Can be Both Uncaused and Up to the Agent. In *Intentionality, Deliberation and Autonomy*, edited by C. Lumer and S. Nannini. Burlington, VT: Ashgate Publishing Company.

Goetz, Stewart. 1988. A Noncausal Theory of Agency. *Philosophy and Phenomenological Research* 49:303-316.

Guevara, Daniel. 2000. *Kant's Theory of Moral Motivation*. Boulder: Westview Press.

Guignon, Charles. 1984. Heidegger's "Authenticity" Revisited. *Review of Metaphysics* (38):321-339.

———. 2000. Philosophy and Authenticity: Heidegger's Search for a Ground of Philosophizing. In *Heidegger, Authenticity, and Modernity*, edited by M. Wrathall and J. Malpas. Cambridge: MIT Press.

———. 2002. Ontological Presuppositions of the Determinism - Free Will Debate. In *Between Chance and Choice*, edited by H. Atmanspacher and R. C. Bishop. Thorverton UK: Imprint Academic.

Haji, Ishtiyaque. 2000. On Responsibility, History and Taking Responsibility. *The Journal of Ethics* 4 (4):392-400.

Harman, Graham. 2010. Technology, Objects and Things in Heidegger. *Cambridge Journal of Economics* 34 (1):17-25.

Heidegger, Martin. 1982. *The Basic Problems of Phenomenology*. Translated by A. Hofstadter. Indianapolis: Indiana University Press.

———. 1992. *History of the Concept of Time*. Translated by T. Kisiel. Indianapolis: Indiana University Press.

———. 1996. *Being and Time*. Translated by J. Stambaugh. Albany: SUNY Press.

———. 1997. *Plato's Sophist*. Translated by R. Rojcewicz and A. Schuwer. Bloomington: Indiana University Press.

———. 2002. *The Essence of Human Freedom*. Translated by T. Sadler. New York: Continuum.

Honderich, Ted. 2002. *How Free Are You?* New York: Oxford University Press.

Hornsby, Jennifer. 2003. Agency and Causal Explanation, edited by A. Mele. New York: Oxford University Press.

Kane, Robert. 1989. Two Kinds of Incompatibilism. *Philosophy and Phenomenological Research* 50:219-254.

———. 1996. *The Significance of Free Will*. New York: Oxford University Press.

———. 2000. Non-Constraining Control and the Threat of Social Conditioning. *The Journal of Ethics* 4 (4):401-403.

———. 2007. Libertarianism. In *Four Views on Free Will*, edited by J. M. Fischer, R. Kane, D. Pereboom and M. Vargas. New York: Oxford University Press.

Korsgaard, Christine. 1996a. Morality as Freedom. In *Creating the Kingdom of Ends*. New York: Cambridge University Press.

———. 1996b. *The Sources of Normativity*. Cambridge: Cambridge University Press.

———. 2008a. *The Constitution of Agency*. New York: Oxford University Press.

———. 2008b. Self-Constitution in the Ethics of Plato and Kant. In *The Constitution of Agency*. New York: Oxford University Press.

———. 2009. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.

Lavin, Douglas. 2004. Practical Reason and the Possibility of Error. *Ethics* 114 (3):424-457.

Levy, Neil. 2005. The Good, the Bad, and the Blameworthy. *Journal of Ethics and Social Philosophy* 1 (2):1-16.

———. 2007. *Neuroethics*. Cambridge: Cambridge University Press.

———. 2008. Restoring Control: Comments on George Sher. *Philosophia* 36 (2):213-221.

Levy, Neil, and Michael McKenna. 2009. Recent Work on Free Will and Moral Responsibility. *Philosophy Compass* 4 (1):96-133.

Locke, John. 1980. *Second Treatise of Government*. Indianapolis: Hackett Publishing.

McDowell, John. 1996. *Mind and World*. Cambridge: Harvard University Press.

———. 2007a. Response to Dreyfus. *Inquiry* 50 (4):366-370.

———. 2007b. What Myth? *Inquiry* 50 (4):338-351.

———. 2009. Conceptual Capacities in Perception. In *Having the World in View*, edited by J. McDowell. Cambridge: Harvard University Press.

McKenna, Michael. 2008. Putting the lie on the control condition for moral responsibility. *Philosophical Studies* (139):29-37.

Nagel, Thomas. 1986. *The View from Nowhere*. New York: Oxford University Press.

Nelkin, Dana. 2000. Two Standpoints and the Belief in Freedom. *Journal of Philosophy* 97:564-576.

Nicholson, Graeme. 2005. The Constitution of Our Being. In *Heidegger's Being and Time: Critical Essays*, edited by R. Polt. Lanham, MD: Rowman & Littlefield Publishers.

Okrent, Mark. 2000a. Intending the Intender (Or, Why Heidegger Isn't Davidson). In *Heidegger, Authenticity, and Modernity*, edited by M. Wrathall and J. Malpas. Cambridge, MA: MIT Press.

———. 2000b. Intentionality, Teleology, and Normativity. In *Appropriating Heidegger*, edited by J. Faulconer and M. Wrathall. NY: Cambridge University Press.

Pippin, Robert. 2007. Can There Be 'Unprincipled Virtue'?: Comments on Nomy Arpaly. *Philosophical Explorations* 10 (3):291-301.

Ricoeur, Paul. 1992. *Oneself as Another*. Translated by K. Blamey. Chicago: University of Chicago Press.

Sadler, Ted. 1996. *Heidegger and Aristotle: The Question of Being*. London: Athlone Press.

Sayre-McCord, Geoffrey. 2001. Criminal Justice and Legal Reparations As an Alternative to Punishment. *Nous-Supplement: Philosophical Issues* 11:502-529.

Scanlon, Thomas. 1988. The Significance of Choice. In *The Tanner Lectures on Human Values*, edited by S. M. McMurrin. Salt Lake City: University of Utah Press.

Searle, John. 2001. *Rationality in Action*. Cambridge, MA: MIT Press.

Sher, George. 2006. Out of Control. *Ethics* 116 (2):285-301.

———. 2008. Who's in Charge Here?: Reply to Neil Levy. *Philosophia* 36 (2):223-226.

Smith, Angela M. 2005. Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics* 115 (2):236-271.

———. 2007. On Being Responsible and Holding Responsible. *Journal of Ethics* 11 (4):465-484.

———. 2008. Control, Responsibility, and Moral Assessment. *Philosophical Studies* 138 (3):367-392.

Staehler, Tanja. 2008. Unambiguous Calling? Authenticity and Ethics in Heidegger's *Being and Time*. *Journal of the British Society for Phenomenology* 39 (3):293-312.

Strawson, Galen. 1994. The Impossibility of Moral Responsibility. *Philosophical Studies* 75 (1-2):5-24.

———. 2000. The Unhelpfulness of Indeterminism. *Philosophy and Phenomenological Research* LX (1):149-155.

———. 2003. Mental Ballistics or the Involuntariness of Spontaneity. *Proceedings of the Aristotelian Society* 103 (3):227-256.

Strawson, Peter F. 1962. Freedom and Resentment. *Proceedings of the British Academy* (48):1-25.

Stroud, Sarah. 2007. Moral Worth and Rationality as Acting on Good Reasons. *Philosophical Studies* (134):449-456.

Talbert, Matthew. 2009. Implanted Desires, Self-Formation and Blame. *Journal of Ethics & Social Philosophy* 3 (2).

Tanzer, Mark. 2001. Heidegger on Freedom and Practical Judgment. *Journal of Philosophical Research* 26.

Taylor, Richard. 1966. *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall.

———. 1983. *Metaphysics*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.

van Inwagen, Peter. 1989. When is the Will Free? *Philosophical Perspectives* 3:399-422.

Vargas, Manuel. 2005. The Trouble With Tracing. *Midwest Studies in Philosophy* XXIX:269-291.

Velleman, J. David. 1992. What Happens When Someone Acts? *Mind* 101 (403):461-481.

———. 2000. Introduction. In *The Possibility of Practical Reason*. New York: Oxford University Press.

———. 2002. Identification and Identity. In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by S. Buss and L. Overton. Cambridge, MA: MIT Press.

———. 2009. *How We Get Along*. New York: Cambridge University Press.

Watson, Gary. 1982. Free Agency. In *Free Will*, edited by G. Watson. New York: Oxford University Press.

Wolf, Susan. 1980. Asymmetrical Freedom. *Journal of Philosophy* 77:151-166.

Zimmerman, David. 2002. Reasons-Responsiveness and Ownership-of-Agency: Fischer and Ravizza's Historicist Theory of Responsibility. *The Journal of Ethics* 6:199-234.

Žižek, Slavoj. 1993. *Tarrying With the Negative*. Durham, NC: Duke University Press.