RESEARCH OVERVIEW

Pamela Hieronymi
September, 2010

I here present an overview of my research. I have attempted, not to summarize each paper, but rather to present the connections between them, the views underlying them, and the concerns that have animated my efforts. In taking this approach, some papers are treated in much more detail than others, sometimes unjustly. Moreover, in the center of the presentation is a sketch of what is still to come—which is, inevitably, under-argued.

INTRODUCTION

Some things we do. Other things just happen to us. It is hard to find a more momentous distinction. The difference between activity and passivity, between being an agent and being a patient, marks (in at least one very important way) our personal boundaries: our possibilities and our limitations, our responsibilities, failures, and achievements, what is up to us or in our control and where we are hostage to luck, fate, or the generosity of others.

And yet this distinction, between between what we do and what we suffer, is very poorly understood. There is no consensus about what marks it—indeed, there is little steady illumination in the area, at all.

Most (though not all) of my research to date can be seen as an attempt to better understand this distinction by attending to what I think of as the interesting middle cases. Start with the extremes: to one side stand intentional actions, like raising your right hand, planting some azaleas, or running for office. These seem clearly to be things you do. At the other extreme are the (relatively) clear cases of passivity. These include not only being blown by the wind, falling towards the earth, aging, succumbing to disease, and winning the lottery, but also having a headache, hearing a ringing in your ears, or seeing the scene before your eyes. These things are not your doing, they are things that you cannot help, did not choose, and are not up to you. Between these extremes stand an interesting class of states of mind, states which appear passive, when contrasted with ordinary intentional actions, but which seem active, when contrasted with sensations or perceptions.

Consider believing. Believing is importantly unlike raising your right hand or running for office, in that you cannot believe "voluntarily" or "at will." While you can raise your right hand or run for office for any reason that you think shows it worth doing—to win a bet, say, or make a joke, or make a point—you cannot believe something (e.g., that the butler did it) in order to win a bet, make a joke, or make a point. You can only believe what you take to be true. You are constrained, in your believing, in a way that you are not, in your action. Thus, when compared to action, believing seems passive.

On the other hand, believing is importantly unlike having a headache or seeing the scene before your eyes. What you believe is up to you. If you think appearances are deceiving, you will not believe your eyes. If, upon considering the evidence, you change your mind about the butler's guilt, you will no longer believe him guilty. Our beliefs reflect or embody our thoughts—in particular, our thoughts about what is so. And thinking is something we do. Thus we, as thinkers, do not simply suffer our beliefs, in the way we suffer a headache or the onslaught of our perceptual experience. Moreover, the assumption that our beliefs are up to us makes sense of our practice of asking others to defend their beliefs, to provide us with their reasons for believing. If we ask someone why she

has a headache or a ringing in her ears, we expect an account of what produced this state of discomfort. But if we ask someone why she believes the butler guilty, we expect to be given, not an explanation of the history of a particular mental episode, but something quite different: a present case for the butler's guilt, a case to support the belief. Such a request makes sense, I contend, only if we are assuming that your beliefs embody your thoughts in a way in which your pains and perceptions do not.

So belief sits in a fertile middle ground between ordinary intentional actions, which seem clearly active, and perception and sensation, which seem clearly passive. I believe many other states of mind—indeed, many of those we find most important—also reside in this middle ground. I would include, here, intentions, decisions, many emotions or attitudes (such as resentment, gratitude, trust, admiration, contempt, or satisfaction at a job well done), and even intentional actions *insofar as* they are characterized or individuated by their motive (so, for example, kind actions, or generous ones, or spiteful ones—planting azaleas in order to cheer your ill neighbor or running for office in order to prove your mettle). And I believe that understanding the sense in which the denizens of this middle ground are active—understanding our agency or control or activity with respect to this class—is the key to understanding both responsible human activity and its limits.

Believing (or Intending) at Will

A natural way to approach this middle class is by considering a standing puzzle about believing at will. It seems more or less plain that, while we can (if able-bodied) raise our right hand or hold our breath at will, we cannot believe that it is raining or that everything will be okay at will. We cannot believe in order to make a joke or to make ourselves feel better. Philosophers have long thought this more than a contingent psychological limitation. It seems, rather, a necessary feature of belief. Yet it has been surprisingly difficult to say exactly why believing is necessarily involuntary.

I address and (I believe) solve this puzzle in an early, foundational article, "**Controlling Attitudes.**" There I show that, though belief is subject to two robust forms of agency, we nonetheless cannot believe at will.

One way in which we exercise agency with respect to our beliefs is by coming to conclusions about what is so and therein forming or revising a belief. This is an example of exercising what I call *evaluative control* with respect to our beliefs. Another way we exercise agency is by taking actions designed to affect our beliefs in certain ways: we can conduct an investigation, or change the world to make something obviously true, or take some medication to quell our anxieties. I call this exercising *managerial* or *manipulative control* over a belief. But, I argue, exercising neither of these forms of agency amounts to believing at will, or voluntarily. In fact, I argue, nothing could count as both a belief and as having been done at will. If I am correct, our inability to believe at will thus represents no shortcoming in our powers: it is an inability to do something that makes no sense—akin to an inability to construct a square circle or to add two by subtracting five.

"Controlling Attitudes" makes the same argument for intention. One way we exercise agency with respect to our intentions is by deciding what to do and another way is by taking actions designed to effect or change our intentions. But neither of these would count as intending "at will," in the sense at issue in the puzzle about believing. Thus, perhaps surprisingly, you can no more intend at will than believe at will.

(In insisting that our agency in believing and intending is largely isomorphic, my work follows a long tradition that understands the will as "reason in its practical employment." In "**The Will as**

Reason," I sketch a way of understanding the difference between the practical and theoretical employments of reason that secures the often overlooked benefits of such a view while avoiding the standard objections to it. Here I address the possibility of weakness of will. In "**Reasons for Action**" I suggest a corresponding, but very minimal, account of the general form that the explanation of action should take. I argue for it by showing how it avoids the difficulties and confusions of more ambitious alternatives.)

In a paper that is nearly completed, "**Believing at Will**" (to appear in *The Canadian Journal of Philosophy* in 2011), I am revisiting these early arguments. Here I focus more on showing how difficult it is to specify the sense of "at will" in the intuitive claim that we cannot believe at will, and I thereby draw out more vividly some of the achievements of my work: I have developed an account of what it is for an activity to be "voluntary" or doable "at will," in the sense at issue, and I have offered a diagnosis of why it seems to us that there is something that we cannot do, some limitation in our powers, even though (if I am correct) there is no such thing. Finally, in revisiting my arguments, I take the opportunity to respond to criticisms of my article (and its consequences) that have appeared in print since its publication.

The Nature of Reasons and the Kinds Thereof

Developing my account of why believing is involuntary required some innovation in thinking about what a reason is. I argue for a modification to the standard account in "**The Wrong Kind of Reason.**" While many philosophers understand a reason to be *a consideration that counts in favor of an action or attitude*, I suggest we do better by relating reasons, not directly to events or psychological states (which seems to me an unkosher blending of the rational and the empirical), but rather, first, to questions.

A reason is an item in a (possible) piece of reasoning, where reasoning is thought directed to a conclusion, i.e., thought directed at a question. So I suggest that a reason is *a consideration that bears or is taken to bear on a question*. (To preview: I take it that actions and attitudes are the events and states of mind that happen as thinkers answer questions. By relating considerations directly to actions and attitudes, the original, unkosher account obscures from view the point at which agency is exercised: in answering a question.)

I argue for my account by showing that it allows us to solve what has come to be called "the wrong kind of reasons problem," a problem I think generated by the alternative account. Philosophers have noted that, for many attitudes (such as belief, intention, desire, preference, valuing) there seem to be reasons of the wrong kind—e.g., the fact that believing it would make you feel better seems the wrong kind of reason for believing that everything will be okay. And yet the fact that it would make you feel better surely counts in favor of believing everything will be okay—just as it counts in favor of getting some rest or taking some aspirin. So philosophers have had difficultly giving a general account of what makes a reason of the right or wrong kind.

The difficulty can solved by adopting my account, because my account forces us to relate reasons to attitudes via questions. Certain attitudes are formed or revised by answering certain questions: by answering for yourself the question of whether everything will be okay, you form or revise a belief that everything will be okay. The right kind of reason for an attitude, then, are those that bear (or are taken to bear) on the question, the settling of which amounts to forming or revising the attitude. I call these the *constitutive reasons* for the attitude. These are the reasons by which one might exercise evaluative control with respect to the attitude.

But, importantly, the class of constitutive reasons does not exhaust the class of considerations that might count in favor of an attitude. A consideration can count in favor of an attitude simply by showing the attitude in some way good to have, without bearing or being taken to bear on the question, the settling of which amounts to forming or revising the attitude. So, the fact that it would make you feel better shows something good about believing everything will be okay, without bearing on the question of whether everything will be okay. I call these remaining reasons the *extrinsic reasons* for an attitude. Extrinsic reasons for an attitude are the wrong kind of reason.

This distinction in kinds of reasons maps onto the earlier distinction in kinds of agency. Constitutive reasons for an attitude, the right kind, are those by which one exercises evaluative control. Extrinsic reasons for an attitude are reasons for which one might exercise managerial or manipulative control over that attitude (extrinsic reasons for a belief that *p* might be constitutive reasons for intending to bring it about that you believe *p*, that is, for intending to manage that belief).

We can now see (as I explain in "**Believing at Will**") that the possibility of encountering the wrong kind of reason leads to the puzzle about believing at will. We are creatures who settle questions and therein form attitudes. And, being reflective creatures, we can think about these attitudes. And, being reflective creatures who live in a world of delights and hazards, we can notice that some of our attitudes are particularly inconvenient, pleasing, embarrassing, useful, or admired. These facts, noticed in reflection, seem to provide us with reasons for or against the attitude. But they are reasons of the wrong kind. And so we will find ourselves in the puzzling position of having reasons for an attitude that is, itself, the kind of attitude adopted for reasons, and yet unable to adopt it for *these* reasons. This can seem a strange limitation of our powers—it seems we cannot adopt the attitude "at will." Thus there is an important connection between the wrong kind of reasons problem and voluntariness: any action or attitude for which we can construct the wrong kind of reason will be one that cannot be done "at will." (This is the large middle class that interests me.)

The Embodiment of Agency, Part One

In "**Two Kinds of Agency**" I present, on its own, the distinction between evaluative and managerial control. I take the usefulness of this distinction to lend support to an assumption (a postulate, perhaps) of all my work: that certain attitudes *embody* our answer to a question or set of questions.

In "Two Kinds," I stipulate that I will there mean, by 'embodiment,' something more complicated but less controversial than what I think is true. In "Two Kinds," to say that a belief that *p* embodies an answer to the question of whether *p* is to assert a conjunction of two relatively uncontroversial conditionals: first, if one settles the question of whether *p*, then one believes *p* and, second, one beliefs *p* if and only if one is committed to a positive answer to the question of whether *p*. This conjunction is assumed in all my work. The simpler but more controversial claim that I think true is this: one believes *p* if and only if one settles positively the question of whether *p*. (And likewise for the other attitudes.)

Fully supporting this simpler claim about embodiment will require **work not yet completed,** though I am approaching it in "**Reflection and Responsibility**." In particular, it will require explaining both how weak the claim is and why, nonetheless, it is both important and necessary. I will here only gesture at both these tasks.

The claim is weaker than it might seem, because, in saying that an attitude embodies the answer to a question, I am *not* suggesting that there is another, independent psychological state or event, the

answering or settling of a question, which we must independently identify and relate to the original attitude. The only psychological state or event at hand is the original attitude. The claim is rather that believing p *is* (or embodies) settling the question of whether *p*; intending to x *is* (or embodies) settling the question of whether to *x*. If asked how we know there is a settling or answering of a question, in a particular case, I will reply, in whatever way we know there is a believing or an intending, in that case. If asked what it is to settle positively the question of whether *p*, I will say, it is to believe *p*. If asked when you settled the question of whether *p*, I will say, you are doing so whenever you believe *p*. (This way of putting things might make it seem that the idea of settling a question is doing none of the work, but that appearance is misleading: we often enough conclude that someone believes or intends by concluding that he or she is committed to an answer to the question of whether *p* or whether to *x*.)

But I now face the question of why this apparently very minimal, deflated, seemingly empty claim about settling or answering is needed; why posit a settling or answering, each time we encounter a believing or intending? Why re-label them thus?

The background answer is that the notion of settling or answering a question is what, I believe, characterizes a genus of which believing, intending, and the entire middle ground that interests me are species; it allows us to see the form and the limits of our agency in this (I believe fundamental) domain. (Recall, it was by relating attitudes to questions that we located the wrong kind of reason and understood the non-voluntariness of any activity for which such reasons can be constructed.)

But a perhaps more helpful answer to "Why re-label them thus?" is that our ordinary ways of thinking about our relation to and responsibility for our beliefs, intentions, and other such attitudes requires that they are active in a way that is captured by the claim that they embody our answering of a question and is not captured by more popular alternatives. Arguing for this answer is one central aim of the rough and ambitious draft, "**Reflection and Responsibility**." Before suggesting (and I will here only suggest) why I think the popular alternatives are inadequate, it will be helpful to see why they are popular. And, for that, it is helpful to consider the traditional problem of free will.

Free Will

I believe that the framework I have developed can be used to diagnose the traditional problem of free will. When thinking about agency, freedom, or control, it entirely natural to focus on the agency we exercise in acting intentionally and the control we enjoy over ordinary objects. We control many things—both our own voluntary actions and ordinary objects—by thinking about them and bringing them to conform to our thoughts. Or, at least, that is how it seems in the paradigm cases: we think about what to do and then do it; we decide how things should be, and then bring it about that they are as we decided. Our ordinary sense of control thus involves both a certain kind of *awareness*—we have in mind what we intend to do—and a certain kind of *voluntariness* or *discretion*—we can decide to do whatever we think worth doing. And so it seems to us that we are in control when, and only when, we are the deliberate cause of our own representations—of that which we represent.

However, we are also reflective creatures, and we can think about ourselves—in particular, we can think about ourselves as the deliberate cause of our own representations. And, when we do so, we will notice that the activity by which we are controlling that which we represent—the representing and bringing about, *itself*—should, itself, be in our control, if we are to be in control of that which we represent. We also need to be in control at this, more fundamental, level.

And now the trouble begins. If we think we are in control when, and only when, we are the deliberate cause of our own representations, then it will seem that we must somehow gain *this* sort of control, now at the more fundamental level. It will seem that we should be able to represent and bring about our own representing and bringing about—more colloquially, to think about and choose our own choosings. If the only notion of control that we have on hand is the one modeled on intentional action and control of ordinary object, then it will seem that, if we cannot think about and choose our own choosings, then we are not, after all, in control of ourselves, and so not, after all, in control of anything.

At this point some become pessimistic. Some of the pessimists see, in our predicament, an immediate and hopeless regress: we will never be in control of ourselves, because each attempt we make to control something will itself involve some activity that was not, itself, controlled. The pessimism of others does not appear until they consider the unfolding of events through time. They reflect on their lives and notice that each decision they make, and each thing they represent and do, can be adequately explained by conditions in place prior to it. And those conditions are not (or, often enough are not) things they control. Whether the larger share goes to nature or to nurture is immaterial—the determinants of their choices were not chosen by them. Worse, their future choices are in the same predicament. And so it seems that have somehow been cut out of their lives; they have no control over it. This, I take it, is the basic threat felt by the garden-variety incompatibilist.

Others, however, are more optimistic. These are the compatibilists. The most popular forms of compatibilism, in recent years, have appealed to hierarchy or self-reflection. Compatibilists of this variety make much of the hopeful fact that, not only can we think about ourselves, but, crucially, thinking about our minds will often enough change our minds: when we reflect upon ourselves, we change ourselves. (On one kind of view, we endorse some aspects of ourselves and reject others, and so incorporate some into while banishing others from our responsible selves. On another kind of view, we inhibit the motivational force of some of our first-order attitudes while aiding the efficacy others. On a third kind of view, our attitudes are sensitive to our higher-order thoughts about their justification.) It is not hard to see why the appeal to self-reflective activity is so attractive: if we can reflect upon and change ourselves, we enjoy a kind of control over ourselves at least very similar to the control exercised in intentional action and over ordinary objects. That is to say, by appealing to self-reflective activity, it seems that we can supply something like the ordinary notion of control to the fundamental case: contrary to the appearance of a regress, it turns out that we can, after all, represent and bring about our own representing and bringing about—or, at least, we can do something close enough to that to secure for ourselves a sense of control over ourselves.

And yet I think the appeal to self-reflection will ultimately prove unsatisfying. My reason for finding it unsatisfying is not, however, the common one.

The common reason to be dissatisfied with the appeal to self-reflection is a sense that it has not addressed the basic threat felt by the incompatibilist. The incompatibilist will consider the self-adjusting, self-sensitive features of our mind to which the compatibilist has drawn our attention and note that their operations, too, are adequately explained by facts that precede them, and not, ultimately, by our own representing and causing. And so the incompatibilist cannot see how adding a more sophisticated layer, which we also do not control, is going to gain control for us. We have added an epicycle, but we have not, thereby, gained a foothold.

The compatibilist often, at this point, starts to ask what kind of control we are looking for, what work the notion of freedom or control does for us. One answer is, we need control if we are to be responsible. And the typical compatibilist then often works to show that the kind of sophisticated, self-adjusting, self-sensitive features of our mind to which he has drawn our notice is sufficient to secure responsibility. Having secured that, he rests content. (The compatibilist typically finds any further notion of control or freedom either mysterious or unnecessary).

But this is just the point at which I am dissatisfied. I do not think that the self-adjusting, self-sensitive features do adequately secure the kind of responsibility we take ourselves to bear for our beliefs and intentions. (I will, in a moment, try to say why.)

I suggest that we could cut through these knots by supplying ourselves with another notion of control. And I believe we already have one. While ordinary action and control of ordinary objects provides one familiar instance of agency, thought provides another. And, in thinking—*whether self-reflectively or not*—one changes one's mind. Thus, I believe, the evaluative control we exercise with respect to our mind as we settle questions, the control we exercise with respect to our thoughts as we think, provides another, also ordinary, notion of agency—though, admittedly, one that is lacking the familiar and comforting features of awareness and discretion. Though it lacks those features, I nonetheless believe that evaluative control supplies us with the fundamental form of agency we exercise as we engage in the more familiar activities of representing and bringing about that which we represent.

(It should be admitted that my alternative also does not *address* the basic threat. Rather, and unlike the appeal to reflection, it rejects the assumption on which the basic threat is based: that we control a thing when and only when we think about it and conform it to our thoughts. I insist that we control our minds even when we are are not thinking about them, even when the changes we effect are not ones we intended to bring about, and even though we cannot effect such changes voluntarily or at will. This may seem odd, but I think it far less odd than it at first appears.)

Two Kinds of Responsibility

Why do I say that the appeal to reflection will not adequately capture the responsibility we take ourselves to bear for ourselves and our attitudes? In the draft "**Reflection and Responsibility**" I attempt to say why. (That draft is still in a preliminary state, so what I say here is also preliminary and as yet inadequate.) I there introduce a distinction between two forms of responsibility and argue that the alternative clearly accounts for only the less fundamental of these two forms—our responsibility for our actions and those things we act upon. By preserving the familiar features of awareness and discretion, the reflective account fails to capture our responsibility for the fundamental activity of deciding or concluding.

Consider first the genus of which the two forms of responsibility will be species. To be *responsible* for something, as I understand it, is to be open to certain sorts of assessment and response on account of that thing. We are thus responsible for our intentional actions: we can be, on account of our intentional actions, open to assessment not only as reasonable or unreasonable, but also as greedy, gracious, petty, courageous, magnanimous, insensitive, and the like, and, so, one can be the appropriate target of certain sorts of reactions, such as resentment, gratitude, admiration, trust, distrust, or esteem.

One (more specific) way to be responsible for a thing is to be (what I call) *answerable* for it. You are answerable for those things for which you can rightly be asked for your reasons. You can, e.g.,

rightly be asked your reasons for intending to sign the contract, believing that the recession is over, being indignant about the war, or sabotaging the mission. You can rightly be asked for your reasons for these things, because they are the kind of thing done or brought about for reasons (whether or not they were done, just now, for reasons). When you are answerable, you are responsible, because, in doing the kind of thing done for reasons (or in holding the kind of attitude for which there are reasons), you adopt or inhabit what I will call an evaluative take on the world and your place in it—you find certain things to be true, worthwhile, or important, while neglecting or rejecting others. Those things for which you are answerable thus reveal what might be called the quality of your will and so open you to the sorts of assessments and responses characteristic of being responsible.

Notice, though, that you can be responsible, in the sense here outlined (rightly open to a certain range of evaluation and response) for a great many things for which you are *not* answerable. You can be responsible for the misbehavior of your dog, the disarray of your apartment, or the functioning of your automobile. If your dog misbehaves, you might be thought negligent or indulgent, and you might be the object of resentment or contempt. But you are not answerable for your dog's misbehavior. You cannot be asked for your reasons for his mischief.

Plausibly, what responsibility you bear for your dog's behavior derives from the fact that you have some hope of controlling your dog— his behavior is something you can affect through your actions —together with the fact that he is, in some sense, yours to control. (You might also have some hope of controlling *my* dog—maybe more hope than I—but she is not yours to control, and so you are not responsible for her behavior.) I say you are responsible for such things because they fall into your *jurisdiction*: they are manageable and in some sense yours to manage.

Notice, too, that those things for which you are answerable *also* fall into your own jurisdiction. Your beliefs and intentions, your resentments and admirations, are things about you that you can and sometimes should take action to change. They are manageable and yours to manage.

Finally, notice that your beliefs and intentions, your resentments and admirations, could *not* reasonably fall into your jurisdiction if you were not able to reflect upon them and somehow bring them into accord with your thought.

Thus, the kind of self-reflective agency appealed to by the alternative account, modeled as it is on ordinary action, can help to ground and explain our *jurisdictional* responsibility for these attitudes: the fact that we can think about and so change these attitudes can explain why we can be expected to ensure that all is well with our actions and attitudes—why we can be expected to ensure that they are well-trained and in good working order, so to speak.

What I do not see is how, or why, the kind of agency that the reflective model imagines explains, conditions, or provides a ground for *answerability*. I try to expose the lacuna in the draft, "Reflection and Responsibility." I there also raise doubts that anything that might be called "reflective control" is a *condition* on answerability. (I also sketch an alternative explanation of the remarkable correlation between responsible creatures and creatures capable of self-reflection.)

In contrast, I think my account of evaluative control readily grounds our answerability. On my account, the attitudes for which we are answerable embody our answers to questions. And, in general, if you answer a question, you can rightly be asked for your reasons for doing so. Further, in answering a question about what is true or good or to be done, you adopt an evaluative take on the world and your place in it. That is, in answering questions, you determine (though an exercise of

evaluative control) the quality of your will. Exercising evaluative control thus opens you to the sorts of assessments and responses characteristic of being responsible.

The Embodiment of Agency, Part Two

It may seem that I have not yet provided an argument for the stronger and more controversial interpretation of 'embodiment.' One might, after all, think that a far more natural account of these attitudes would claim that often, though not always, we *form* or *revise* them by answering for ourselves a question, and that, having been formed—whether actively or not—they are simply dispositional states for which we are answerable, due to the fact that we *could* form or revise them by reconsidering a question. That is, they *commit* us to an answer to a question, but they do not embody the answering of a question, in the stronger sense—they only reveal that a question might once have been and might yet be answered.

In reply I need, again, to show both how weak the claim is and why, nonetheless, it is both important and necessary. For the second of these, I will attempt to show that answerability (and the relevant sense of commitment) presumes present, not past and not merely possible, activity. When we take you to be answerable for your beliefs or intentions, we do not treat your beliefs or intentions like the paper you wrote last week or like the order you issued to your staff yesterday—as the standing, as-yet unrevised product of some earlier thinking. Nor do we treat your beliefs or intentions like the wine you accidentally split on your host's carpet—something that might have been, but in this case was not, the product of your agency, but which nonetheless saddles you with certain commitments. When you ask me why I believe the butler did it, you are not asking me why I once formed that belief and stored it in memory; you are asking me, now, for a present case for the butler's guilt. My answer is no less an answer to your question if I come up with new reasons, on the spot. And if I reply that I no longer take the butler to be guilty, I would, thereby, reject your question. By such reflections, I hope to show it misleading to model attitudes merely as dispositional states that we might find ourselves with, which we can also form or affect, though episodes of activity, at various moments in time.

Such reflections may seem terribly subtle threads on which to hang the claim that a state of mind embodies an activity—a claim that smacks of a category mistake. But, again, the claim is weaker than it might at first appear. It will seem implausibly strong if you bring to it an independent account of what activity must be like: if you think an activity must be, for example, a process of change that unfolds through time (or that it must have an aim, or that it must involve motion). But I can, and do, deny that the kind of activity at issue involves a process of change (and that it has an aim or involves motion). Rather, I mean to claim that believing, intending, and the like are activities in whatever sense that they need to be, to make sense of our ordinary ways of thinking—to make sense of the demands and standards to which they are subject. I mean to show that these demands and standards do not make sense if we think of these as states to which we are passive, states that we can merely affect through episodes of activity. So, we need a notion of activity suited to them. But given that the notion of activity at issue is simply dedicated to answerability, I do not think it at all costly to insist that certain states of mind embody activity—that they are its psychological face, so to speak. Rather, I think, all the costs are incurred in denying this and then attempting to use some *other* notion of activity (a process of change, or an ordinary action, or a special kind of efficacious self-reflection) to model the kind of activity that we do, in fact, presume is necessary to ground responsibility.

These last few sections have forayed into the leading, and very rough, edge of my work. I present them hoping they help to situate the work I have completed. I now return to the completed work.

Responsibility, Blame, Expectation, and Obligation: A Moral Theory to Fit

My particular variety of compatibilism is, by one measure, quite extreme. It obviously would not fit together with certain conceptions of responsibility, blame, or moral demand. In particular, it will not fit with any conception freighted with incompatible assumptions about agency. So, e.g., if one thinks that to be responsible is to be capable of choosing one's future independently of one's past, then obviously I cannot argue that such responsibility is compatible with a recognition that we are not independent of our past. Or, if one thinks that blaming someone includes charging that she neglected an opportunity to have done differently—if you think that, in blaming someone, you are saying, in effect, "You did something wrong, and, moreover, you had an opportunity to avoid this error and, nonetheless, you chose wrongdoing"—then it will not seem that the kind of agency I have sketched will make blame appropriate. It makes no mention of opportunities.

I am not troubled that these conceptions of blame or moral demand will not be supported by the account of agency I have provided, because I find these conceptions unattractive on their own terms. In a series of papers I put forward the alternative view of responsibility, blame, and moral demand that I find more attractive.

The broadest overviews of my approach to these topics are found in two somewhat polemical papers that address the work of contemporary philosophers.

In "**Making a Difference**," I take issue with John Martin Fischer's semi-compatibilism; in particular, I claim that Fischer has conceded too much when he grants that the truth of determinism would show that we do not make a difference. In the process, I provide a sketch of (my take on) the dispute between compatibilists and incompatibilists and suggest that some of the discussion has confused the freedom required for moral responsibility with a very different notion of autonomy, one which is, instead, required to make a certain sort of important difference—a kind of difference that some people surely do make.

Perhaps the broadest and most intuitive account of my picture of moral demand and moral responsibility appears in a paper addressing the view of Michael Smith, "**Rational Capacity as a Condition on Blame.**" In this paper I argue that moral demands, like most other demands, do not bend to accommodate the moral abilities of the individuals to whom they apply. They are, rather, one-size-fits-all. And so, given the vagaries of human life and the hazards of moral development, they will fit some very poorly. This is a tragic fact, but it should not be surprising. If moral demands were to recede in the face of moral inability, if they were to bend to the find the moral capacity of each individual to whom they apply, they could not do the job of adjudicating the competing interests of those sharing a world of limited resources with others equally real. (The view of moral demands here presented could be elaborated upon with the contractualism of "Of Metaethics and Motivation.") Moreover, insensitivity to individual ability is the norm for most of the demands we encounter in life. (Pedagogical demands are properly custom-fit to the individual—to what the student or child can just barely do. But moral demands are not pedagogical.) The demands of parenting or being president, e.g., do not ease just because a given parent or president is too selfish or too stubborn to satisfy them. Rather, we simply hope that the parent or the president will, under the pressure of the demand, improve. Sometimes they do. Sometimes they do not. When we fail to meet a demand, whether through negligence, inattention, or inability, we (and

others) will suffer whatever the consequences of that failure turn out to be—in the case of moral failure, the consequence will be, at the minimum, substandard interpersonal relationships.

The thought that moral demand should accommodate itself to the moral ability of the individuals to whom it applies draws life from a thought often summarized as "ought implies can." Although, as seen above, I am a staunch advocate of the thought that responsibility (in particular, answerability) implies activity, I am in strong disagreement with the typical employments of the slogan "ought implies can." The slogan is typically employed in the contrapositive: a lack of ability implies a lack of obligation. And, in fact, the typical thought is not really about obligation, but rather about blame: if someone lacks the ability to avoid blame, then blaming her is inappropriate. And the motivation for *this* restriction is typically a sense of fairness: blame is a serious matter, and being blamed is a bad thing, and it is not fair to impose that burden on someone who had no opportunity to avoid it.

While the issues here are difficult, I address a core aspect of this thought in "**The Force and Fairness of Blame.**" There I deny that blaming judgments can be rendered unfair by the fact that they are burdensome for their target. To put it in terms of the work earlier described: the fact that being blamed is bad for the blamee is the wrong kind of reason to avoid the blaming judgments that create much of that burden.

Towards the end of "The Force and Fairness" I attempt to compare and contrast the account I present with the strategy famously pursued by P. F. Strawson. In particular, I try to draw out the under-appreciated sense in which the reactive attitudes are *reactive*—they are not voluntary, in the sense of voluntariness I have outlined, and this has very important implications for their justification.

Strawson's work receives further attention in the final section of "**Sher's Defense of Blame,**" where I argue that George Sher's attempt to show that a commitment to blame is entailed by a commitment to morality fails. I suggest that we might find the kind of entailment Sher hopes for by following Strawson more closely.

In "**Responsibility for Believing**" I argue against a different variation on "ought implies can": the claim that responsibility implies voluntariness. As will be clear by now, though I take responsibility to imply activity, I do not take it to imply voluntariness. In fact, I have argued that that for which we are most fundamentally responsible—the quality of our will, or our evaluative take on the world and our place in it—could not be voluntary.

In taking a broadly Strawsonian approach to blame and responsibility, I share the sensibility of T. M. Scanlon. In "**Of Metaethics and Motivation: The Appeal of Contractualism**" I try to present Scanlon's view of moral obligation and permissibility, a view which fits nicely with the account of responsibility I have offered. In particular, I try to present Scanlon's appealing answer to what he once called "the question of motivation" and the relation of this answer to the more metaethical "question of subject matter." I then defend Scanlon's view against various, standard objections, which, I claim, simply misunderstand it. I close by considering what it would take to wed Scanlon's attractive answer to the question of motivation to another, non-contractualist, theory. I conclude that, even if the marriage could be arranged, a good part of the appeal of contractualism would inevitably be lost. In particular, we would lose the place secured, in Scanlon's contractualism, for freedom of conscience. (This is a large piece of work and represents a newer direction for my research, but one whose connection to the rest is, I hope, clear enough.)

My earliest published paper, "**Articulating An Uncompromising Forgiveness**," addresses a topic to which I have not returned, but anticipates many of the positions I have since adopted.

In "**The Reasons of Trust**" I address another seemingly free-standing topic. In fact, however, trust serves for me as a test-case: it is a simpler case in which to think about an argument I hope to construct for the more complicated case of virtuous (and vicious) action. (This target argument appears in a draft titled "Extrinsic Reasons, Alienation, and Moral Theory.")

I assume that an action is, e.g., kind (rather than, e.g., conscientious or spiteful) if it was performed for certain sorts of reasons. If you decide to help your colleague because she is exhausted, then your helping is, presumably, kind. If instead you help her in order to ensure the job is done competently, your helping is conscientious—or perhaps meddling. Which adjective describes your action depends on the reasons for which you acted. The reasons that would qualify your action as, e.g, kind I call *reasons constitutive of kindness*.

We can, then, construct a wrong kind of reasons problem for virtuous action: the reasons constitutive of kindness do not exhaust the considerations that count in favor of performing a kind action. You might, e.g., want to impress your colleagues by acting kindly, in order to advance your own self-serving agenda. This prudential reason counts in favor of performing a kind action, but it is not a reason constitutive of kindness. It is an *extrinsic reason for performing a kind helping action*.

Though it is not entirely straightforward, I believe I can argue that, just as you cannot believe at will, so you cannot perform a kind (or conscientious or spiteful) action for reasons extrinsic to it. Insofar as you act on the extrinsic reasons, to that extent you will *not* be doing what the reason recommends. You will be, at best, engaging in self-management, acting so as to bring it about that you perform a kind action. If I can show that you cannot "directly" perform an action for reasons extrinsic to it, I can then argue that, insofar as moral theories hope to justify moral action, they must do so by providing reasons that are constitutive of the actions they require. (I can also diagnosis a good bit of relatively recent dissatisfaction with moral theory, coming from the likes of Bernard Williams.)

In "**The Reasons of Trust**" I present a roughly analogous argument, for the case of trusting someone to do something. I argue that trust requires what I call a trusting belief. I understand a trusting belief as analogous to a kind action: a belief is trusting if it is supported by certain reasons. I suggest that the reasons constitutive of a trusting belief concern the trustworthiness of the one trusted, rather than the importance of a trusting response. I then argue to a conclusion I find at once surprising and intuitive: to the extent that your reasons for doing what you do concern the importance of a trusting response, to that extent you do something other than trust (you act so as to encourage or promote trust, or to build the esteem of the other, or to discharge duties of trust, etc.). Thus, the degree to which you trust varies inversely with the degree to which you must rely on reasons that make explicit reference to the value of trusting.