Washington University in St. Louis
Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations

Arts & Sciences

Spring 5-15-2015

Paradigmatic Self-Deception

David Stephen Winchell
Washington University in St. Louis

Recommended Citation

Winchell, David Stephen, "Paradigmatic Self-Deception" (2015). *Arts & Sciences Electronic Theses and Dissertations*. 459.
https://openscholarship.wustl.edu/art_sci_etds/459

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Philosophy

Dissertation Examination Committee:
John Doris, Chair
Eric Brown
Dan Haybron
John Heil
Ron Mallon

Paradigmatic Self-Deception

by
David S. Winchell

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2015
Saint Louis, Missouri

Table of Contents

Acknowledgements

This dissertation would not have been possible without the intellectual, emotional, and financial support of many family members, friends, colleagues, and advisors. First, I'd like to thank my brother Jeff and my sisters Margaret, Suzanne, and Laura for all the laughter and conversation that has sustained me over these many years of graduate school. And I would be nowhere without my parents, Mark and Melinda, who have never been less than totally supportive of my urge to follow my intellectual curiosity wherever it leads me, even when that isn't in an especially practical direction, and who have loved me unconditionally.

Thanks too to my friends past and present at the restaurant Niche, where working in the kitchen provided an invaluable creative outlet and taught me an enduring respect for hard work of a very different kind than got me to my PhD: Kayla Kidd, Sarah Osborn, Josh Poletti, Brian Coltrain, Matt Daughaday, Brian Moxey, Todd Stringer, Brian Mason, Clare Wallace, Christopher Kelling, Jennifer Masur, Samm McCulloch, Kyle Mathis, Henry Mitchell, Carrie Gilbertson, Alec Schingel, Alex Feldmeier, Michael Pastor, Tom McGill, Adam Altnether, Nate Hereford, and especially Gerard Craft for the opportunity to work there at all. Outside of Niche, Maggie Walker, Matt Evans, Jonathan Coveney, Annie Winter, Erika King, Eric Pan, and Mary Meyer have all made these last six years immeasurably richer.

My colleagues at Washington University and elsewhere have proven themselves to be not only brilliant but also tons of fun both on and off campus: Trey Boone, Serife Tekin, Felipe Romero, Nazim Keven, Lauren Olin, Gary Williams, Isaac Wiegman, Jason Gardner, Nate Adams, Nich Baima, Tyler Paytas, David Speetzen, Martin Turner, Santiago Amaya, Jeff Dauer, Christiane Merritt, Ian Tully, Kate Shrumm, Mark Povich, Cameron Evans, and Dylan Doherty have all contributed directly or indirectly to the success of this project.

I first discovered my love of philosophy as an undergraduate at the University of Pennsylvania, where I took fantastic courses with Paul Guyer, Susan Meyer, Cristina Bicchieri, Giacomo Sillari, Susan Schneider, Elisabeth Camp, Stephen Perry, and Leo Katz. Special thanks to my first philosophy TA, Jason Rheins, whose passion was infectious, and especially to Adrienne Martin and Scott Weinstein, who guided my intellectual and personal development as a scholar far beyond my expectations, both in and out of the classroom. I am also deeply grateful for the financial support of Washington University, whose generosity via fellowships and teaching assistantships has kept me comfortably fed and housed over the course of my degree. My final thanks go to the faculty, past and present, of the Department of Philosophy and the Philosophy-Neuroscience-Psychology Program, whose dedication and overall excellence continue to inspire me. I have received exceptional mentorship from: Dennis DesChene, Mark Rollins, Claude Evans, Julia Driver, Anne Margaret Baxley, Charlie Kurth, Frederick Eberhardt, Gillian Russell, Mariska Leunissen, Carl Craver, Simine Vazire (Psychology), Roy Sorensen, Lizzie Schechter, Kit Wellman, Dan Haybron (SLU), and especially Ron Mallon, John Heil, and Eric Brown.

John Doris, my dissertation advisor, deserves all the praise I can give him both for his outstanding commitment to my philosophical development and for his kindness and insight on matters both personal and professional throughout this emotionally and intellectually grueling process. Having heard John's urging to "dig deep" in my work more times than I can count, I hope that with this dissertation I've finally gotten to the bottom of something.

David Winchell

Washington University in St. Louis

May 2015

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

Paradigmatic Self-Deception

by

David S. Winchell

Doctor of Philosophy in Philosophy

Washington University in St. Louis, 2015

Professor John M. Doris, Chair

I present a theory of what I call "paradigmatic self-deception," the most serious and vivid cases of self-deception. While there has been much philosophical discussion of self-deception in recent decades, existing work does not explain how and why some cases of self-deception are more severe than others, when self-deception is intentional, and when an individual may justifiably be held morally responsible for being self-deceived. This dissertation answers all three of these important questions. The first chapter reviews the existing literature and motivates the need for a theory of paradigmatic cases. In the second chapter, I introduce three characteristic features of paradigmatic cases as revealed in the Mitt Romney campaign's self-deception during the 2012 presidential election. Next, the third chapter explores how paradigmatic cases of self-deception are plausibly intentional, drawing on the empirical literature on attention and mindfulness meditation to argue that in paradigmatic cases self-deceivers make sophisticated use of attentional capacities to sustain their false beliefs and that the intentions under which these capacities are deployed are *relevant* to self-deception in a way that makes the overall activity of self-deception intentional. Finally, in the last chapter I claim that paradigmatic cases involve culpable ignorance, which I reduce to a kind of negligence: because the paradigmatically self-deceived are negligent, we can hold them morally responsible.

Chapter I
In Search of a Good Theory

Here's a deep fact about human beings: we lie to each other. Sometimes we lie to protect the people we care about, as when parents keep the magic of the holidays alive by letting their kids believe that Santa Claus really does fly around the whole world in one night and really did read their wish lists and leave all those brightly wrapped packages under the tree. We also often lie to make ourselves look better. We exaggerate our successes and gloss over our failures and flaws to enhance our reputations and create idealized versions of ourselves in the minds of others. All of this we do sometimes for good reasons, others for no reason at all, and still other times with malice. And depending on where you fall on debates in moral philosophy, deception can often be morally acceptable and even morally required.

But can we also lie to *ourselves*? That is, can we create and sustain false beliefs in ourselves in ways that further our own interests and desires? If so, that is *self*-deception. The ability to readily self-deceive could come in mighty handy: if I realize I would be happier believing my dismal performance on last week's math exam was due to simple lack of sleep rather than a serious lack of preparation, I could use my powers of self-deception to replace that clear-eyed realization with a comfortable fantasy. But of course self-deception cannot be so straightforward. We cannot edit our beliefs with the same sort of voluntary control we use to shop for groceries, roaming the aisles and adding and removing items as we like. Instead, our control over self-deception must be more subtle and indirect.

But wait! At first blush self-deception might appear to be straightforwardly impossible: no matter how much I may want to, I cannot return to my childlike innocence about Christmas simply by trying hard to resurrect the false belief that Santa Claus exists. Surely, since I have in mind the very belief whose opposite I wish to deceive myself about, self-deception must undermine itself! This and other puzzles all must be overcome in the course of explaining self-deception.

There has been so much discussion of self-deception in recent years that it may seem as if there is little room for a major new contribution. But I fear that, in their zeal to unravel the conceptual tangles of self-deception, philosophers have often ignored the importance of studying the phenomenon as it exists in the real world: there are actual people all around us who are self-deceived about many important things, and a philosophical analysis of self-deception can and must draw on the details of these cases to see what general features they share. Self-deception presents not just fascinating conceptual questions but critical empirical ones: what are the actual psychological and cognitive means by which self-deception is initiated and sustained? Questions like these are at least as important as those that have traditionally occupied philosophers.

1.1 The Project

This dissertation presents a new analysis of self-deception focusing on what I call *paradigmatic* cases. In my view, paradigmatic cases are the most serious and most interesting instances of the general phenomenon. The next chapter explores the features of paradigmatic cases; I argue that these cases tend to be diachronic and involve false beliefs that are both sharply at odds with reality and require significant effort to sustain. In the third and fourth chapters, I explain how paradigmatic cases involve *intentional* self-deception and then that this lets us hold self-deceivers morally responsible in those cases.

But before we get to all of that, I need to lay some groundwork. I contend that existing accounts of self-deception are incomplete and that a theory of paradigmatic cases is necessary for a "full" understanding of the phenomenon. First, then, I need to provide a survey of what the existing accounts look like and describe how they are unsatisfying. My project in this chapter is thus twofold: I will (1) survey the literature on self-deception argue that Alfred Mele's theory is the best of the current accounts, and (2) describe three conditions that a full theory of self-deception – mine – will satisfy and discuss why those conditions matter. In particular, I claim that fully characterizing self-deception requires understanding it (A) as centrally involving "bad faith" in Jean-Paul Sartre's sense, (B) via engagement with relevant empirical literature from cognitive science and psychology, and (C) in terms of substantial real-world cases rather than fanciful thought experiments. Mele's theory satisfies only one of these (B), and so it leaves unexplained much of what is puzzling about self-deception. Let's consider each of these three conditions in turn.

1.2 Condition (A): Bad Faith

In his famous 1943 treatise *Being and Nothingness*, Jean-Paul Sartre devotes the first chapter to an exploration of the concept of *mauvaise foi,*or "bad faith," which he describes as a kind of "self-negation." Part of bad faith involves our capacity to lie. A person who lies holds the truth in mind while knowingly presenting a fiction to someone else:

> The essence of the lie implies in fact that the liar is actually in complete possession of the truth which he is hiding. A man does not lie about what he is ignorant of; he does not lie when he spreads an error of which he himself is the dupe; he does not lie when he is mistaken. The ideal description of the liar would be cynical consciousness, affirming truth within himself, denying it with his words, and denying that negation as such (p. 71).

While "cynical consciousness" is uncontroversially part of interpersonal lying, Sartre boldly claims that in bad faith "what changes everything is that in bad faith it is from myself that I am hiding the truth. Thus the duality of the deceived and the deceiver does not exist here" (p. 72). So bad faith ends up being a form of self-deception: the person who acts in bad faith is able to lie to herself and somehow know the truth even as she works to convince herself otherwise. Bad faith is immediately useful in helping show what self-deception is *not*: it is not simple ignorance, or making a mistake, or convincing yourself of something that you could not reasonably be expected to have known was false. In all of those cases you could cause yourself to believe something that is false, but that is not bad faith and not self-deception because you have in no sense *deceived* yourself.

Assuming we accept that literally lying to yourself is part of bad faith and self-deception, Sartre still owes an explanation of how psychologically convince ourselves of false things. To get there, Sartre emphasizes the "faith" part of bad faith:

> Bad faith apprehends evidence but it is resigned in advance to not being fulfilled by this evidence, to not being persuaded and transformed into good faith. It makes itself humble and modest; it is not ignorant, it says, that faith is a decision and that after each intuition, it must decide and *will what it is*. Thus bad faith in its primitive project and in its coming to the world decides on the exact nature of its requirements. It stands forth in the firm resolution *not to demand too much*, to count itself satisfied when it is barely persuaded, to force itself in decisions to adhere to uncertain truths (p. 91).

That is, Sartre thinks of bad faith as a kind of *attitude* towards the world, and in particular towards the evidence a person encounters that ordinarily has the power to shape his beliefs about the world. In bad faith, the person resolves not to be too demanding towards the evidence he encounters; his evidentiary standards are relaxed compared to someone not in bad faith. Given this stable attitude towards evidence, acting in bad faith should thus turn out to be a simple matter of allowing the attitude to guide his response to any relevant evidence he encounters.[1]

As we will see in later chapters, I am broadly sympathetic to Sartre's account of bad faith as a theory of (paradigmatic) self-deception. I too believe that self-deception in paradigmatic cases involves a certain kind of stable policy of handling evidence in ways that perpetuate a false belief, and that policy can fairly be described as being "undemanding" of evidence such that the state of being deceived persists despite evidence that would shake the confidence of a "normal" person. And I further believe that the most interesting – the paradigmatic – cases of self-deception are instances of Sartrean bad faith. Sartre unpacks his concept of bad faith through a straightforward examination of the nature of deception as applied intrapersonally. Any theory of self-deception that rejects bad faith owes a serious alternative explanation of how that theory really is a theory of deception, and a theory that rejects a straightforward understanding of deception also loses significant plausibility against any theory that *can* capture the intuitive notion of deception; by going along with Sartre, a theory of self-deception is already on solid intuitive footing.

So even writing early on in the history of the literature on self-deception, Sartre managed to get quite a lot right. But his discussion lacks the sophistication found in more recent accounts that have benefited from the lively philosophical debate on self-deception that has raged in recent decades. We must get clearer on the conceptual foundations of self-deception and especially on the empirical work in cognitive science and psychology that gets us beyond *a priori* just-so stories about how self-deception operates. Let's turn now to look at some existing accounts of self-deception. Following Sartre and applying the interpersonal model of

[1] Depending on the situations in which the self-deceived person finds herself, maintaining bad faith could turn out to be easy or incredibly difficult. This is exactly what my theory of paradigmatic self-deception predicts.

deception to self-deception creates two serious conceptual puzzles, so it is important to find an account that solves the puzzles while preserving Sartre's "bad faith" intuition if at all possible.

1.2.1 The Core of Self-Deception

Accounts of self-deception vary considerably, but there are certain elements that everyone tends to agree are part of the phenomenon. Just as in other-deception, self-deception involves a false proposition – the thing that you are self-deceived *about*.[2] The content of the proposition doesn't much matter; you can be self-deceived about earth-shattering things and about trivial things, about things related to yourself and about things related to others or features of the world. What matters is simply that you are deceived about it – though what that means is up for debate.

On most accounts, being self-deceived means actually believing the false proposition, though some philosophers drop this requirement. Many accounts also feature a belief in that proposition's opposite, which is the truth you are in some sense "hiding" from by deceiving yourself. All of this is controversial, as we are about to see. Further debate revolves around *intentionality*: can I intentionally deceive myself, and, if so, are intentional cases central to the phenomenon? While intentionality is a central part of other-deception – if I cause you to believe something I know to be false only accidentally, that doesn't count as deception – there is considerable disagreement about the role of intention in self-deception. In the next sections,

[2] I won't discuss this in the dissertation, but I allow that self-deception need not always involve a *false* belief. All you really need is a belief that is at odds with the evidence that is reasonably available. Suppose, for instance, that L. Ron Hubbard's "revelations" about humans' descending from aliens happen to be true. That is, they are *accidentally* true. If self-deception requires a false belief it follows that Hubbard could not have been self-deceived about our species' ancestry. But because there is no reasonably available evidence that his belief was true – on no plausible epistemological theory is revelation considered a guide to truth, and we do not have any way of obtaining direct knowledge about the past – then we can preserve the intuition that Hubbard was self-deceived. There is probably nothing simple to say about when evidence is reasonably available, so I will not leave this issue aside. It should not affect any of the discussion to follow.

I will introduce two important theoretical puzzles about self-deception and discuss several theories of self-deception that might solve them. Throughout, I will keep Sartre well in mind: an ideal account of self-deception will solve the puzzles while keeping the spirit of Sartre's view alive.

1.2.2 The Static Puzzle

The most natural approach to self-deception is to model it on other-deception, just as Sartre did in his discussion of bad faith. But doing so turns out to raise serious conceptual problems: is self-deception is possible at all on this model? The rough skeptical argument, as sketched by Brian McLaughlin (1988), is that self-deception requires *self-induced deception*, which is impossible. Why think this? One reason comes from what Alfred Mele (1997, 2001) has called the *static puzzle*, the idea that self-deception requires an impossible state of mind:

> The lexical approach is favored by theorists who deny that self-deception is possible. A pair of lexical assumptions are common:
> 1. By definition, person *A* deceives person *B* (where *B* may or may not be the same person as *A*) into believing that *p* only if *A* knows, or at least believes truly, that ~*p* (i.e., that *p* is false) and causes *B* to believe that *p*.
>
> 2. By definition, deceiving is an intentional activity: nonintentional deceiving is conceptually impossible.
>
> [...]If assumption 1 is true, then deceiving oneself into believing that *p* requires that one know, or at least believe truly, that ~*p* and cause oneself to believe that *p*. At the very least, one starts out believing that ~*p* and then somehow gets oneself to believe that *p*. Some theorists take this to entail that, at some time, self-deceivers both believe that *p* and believe that ~*p*. And, it is claimed, this is not a possible state of mind: the very nature of belief precludes one's simultaneously believing that *p* is true and believing that *p* is false (Mele 1997, p. 92).

McLaughlin argues that the static puzzle disappears once we take a more sophisticated approach to deceit. One way around the puzzle is what he calls "a memory-exploiting stratagem" (1988, p. 31). Suppose Mary knows she has a poor memory and wants to miss an unpleasant meeting in several weeks. So she writes down the wrong date for the meeting on her calendar and knows she will forget having done this. Sure enough, she looks at her calendar several weeks later and, trusting its reliability, acquires the false belief that the meeting is at a different time. Her "stratagem" has paid off.

McLaughlin worries that these stratagems may not explain how it is possible to deceive our "contemporary" selves, not just our future selves; after all, self-deception is supposed to be possible synchronically – not just diachronically – if we are truly to defuse the static puzzle. But, he argues, thinking of the Mary case as one in which she merely deceives her future self is to play fast and loose with the notion of temporality. Mary *is* deceived, and it doesn't much matter whether the deceit was perpetrated by an earlier "version" of her; if she were to suddenly remember writing down the wrong date on her calendar, there would still be a case of self-deception to explain; self-deception would still be possible. Of course, if self-deception is dependent on forgetfulness then it becomes a vastly less interesting phenomenon: only incredibly forgetful people are able to deceive themselves, which goes against the notion that self-deception is a powerful, pervasive phenomenon in daily life. Surely vanishingly few people can deceive themselves in the manner of Mary.

But McLaughlin is convinced that holding contradictory beliefs simultaneously is possible – Mary can believe that the meeting is at the wrong time while also believing it is at the correct time – so he thinks we don't need to say that forgetting is required to solve the static puzzle. *Is* holding contradictory beliefs possible? Many philosophers think so, and if they're right then the static puzzle isn't a problem at all. Perhaps most famous is Donald Davidson (1987), who claims that self-deception arises due to an explicit intention to form a belief in the opposite of some proposition the individual knows to be true but wishes were false. This belief cannot be formed directly, of course; instead, the individual "[must] *do* something with the aim of changing his own views" (p. 87). "Doing something" takes the form of self-manipulation, such as shifting inconvenient evidence "into the background." Davidson gives the case of Carlos, who believes he will fail his upcoming driving test but hates the idea of failing and so resolves to believe that he will not. Thus Carlos acts "in such a way as to cause himself to believe" that he will not fail (p. 88).[3] Ariela Lazar (1998) interprets Davidson as arguing that Carlos' false belief is "a consequence of practical reasoning: it is viewed as a consequence of an attempt to fulfill a non-truth-oriented goal" (p. 22).

And since self-deception begins for Davidson with a belief in the true proposition P coupled with a desire to believe not-P, self-deception ends up incorporating both beliefs. Davidson argues:

> [T]he state that motivates self-deception and the state it produces coexist; in the strongest case, the belief that p not only causes a belief in the negation of p but also sustains it. [C]ore cases of self-deception demand that Carlos remain aware that his evidence favors his belief that he will fail, for it is awareness of this fact that motivates his efforts to rid himself of the fear that he will fail (pp. 89-90).

So Davidson thinks that both beliefs have a crucial role to play in sustaining self-deception; the deception would collapse if one belief were lost. To explain how holding two contradictory beliefs works, he appeals to the notion of a "divided" mind, in which mental states in certain parts of the mind may be inaccessible to introspection. I believe such a model of the mind is eminently plausible and take up the issue in detail in Chapter III. For now, it is sufficient to note that Davidson's theory of the mind and of self-deception in general cannot readi-

[3] Davidson is coy about the details of how self-manipulation is implemented so his proposal is hard to take seriously on empirical grounds. But in Chapter III I discuss how attention can be directed so as to facilitate self-deception, and this activity can be construed as acting to cause ourselves to (continue to) believe false things. Unlike Davidson, though, I do not think we deceive ourselves using such explicit intentions; "resolving to believe" that he will not fail the driving test is likely to be self-undermining for Carlos. Davidson's discussion is useful for getting a "two-beliefs" model of self-deception on the table, however.

ly be made ridiculous; the static puzzle is indeed not as troubling as it seemed. And note that Davidson also solves the static puzzle while leaving room for bad faith: it is only because the self-deceived person holds the attitude of bad faith towards relevant evidence that she is able to maintain her self-deceived state.

1.2.3 The Dynamic Puzzle

Mele (1997) also offers a second, *dynamic* puzzle:

> In general, A cannot successfully employ a deceptive strategy against B if B knows A's intention and plan. This seems plausible as well when A and B are the same person. A potential self-deceiver's knowledge of his intention and strategy would seem typically to render them ineffective. On the other hand, the suggestion that self-deceivers typically successfully execute their self-deceptive strategies *without* knowing what they are up to may seem absurd; for an agent's effective execution of his plans seems generally to depend on his cognizance of them and their goals. So how, in general, can an agent deceive himself by employing a self-deceptive strategy (p. 10)?

So where the static puzzle made trouble for the possibility of being self-deceived *in the moment*, the dynamic puzzle raises problems for self-deception as an activity that unfolds over time: forming an intention to self-deceive and then acting on it. Is self-deception "a logically and psychologically possible process," in Mele's words (1987, p. 10)?

Davidson's view, which looked like a promising solution to the static puzzle, runs in to serious trouble here. Carlos is supposed to be able to deceive himself by "resolving" to believe that he will pass his driving test, but Mele persuasively illustrates how doing so seems borderline impossible. That is, Carlos knows that he is likely to fail the test, and having that in mind would surely render ineffective any intention to form the contrary belief. Davidson might appeal to divisions in the mind to argue that the knowledge of Carlos's likely failure can be hidden away from conscious awareness. But since he appears to say that Carlos's resolution to deceive himself is *explicit* – Carlos tries to act on an intention with content "FORM THE BELIEF

THAT I WILL PASS MY DRIVING TEST" – it is hard to see how that intention will not inevitably and continuously bring that knowledge to mind, frustrating the attempt at self-deception. So Davidson's model of self-deception will not get us past the dynamic puzzle – we need something else that can handle both puzzles.

Alternatively, the skeptical response here is to say that the dynamic puzzle rules out self-deception altogether, at least as traditionally conceived. McLaughlin (1988) proposes that the puzzle could force a retreat to the claim that self-deception is just a matter of "insincere thought:" the self-deceived person thinks about something that is false, something he wants to convince himself of, and is then "taken in by it" (p. 39). This is not what we tend to think of when we talk about self-deception, but perhaps the puzzle doesn't allow any other solutions.

Yet, as McLaughlin notes, an reducing self-deception to insincere thought is an over-intellectualization of how self-deception operates. Theories of self-deception vary on how the deception originates and is sustained, but they tend to share the understanding that self-deception is a feature of a human life, one full of activity. We do not deceive ourselves all at once, home alone in the armchair, but instead continuously and even repeatedly throughout our daily lives as we interact with each other and the world. In that context, viewing self-deception as the product of mere insincere thought is a radical oversimplification; indeed we have already seen one potential response in Davidson's view.

An alternative, sophisticated approach to defusing both the static and dynamic puzzles comes in the work of Robert Audi (1982), who claims that the self-deceived individual need *not* hold a false belief He asserts that:

> S is in self-deception with respect to p if and only if
> (1) S unconsciously knows that not-p (or has reason to believe, and unconsciously and truly believes, not-p);

(2) S sincerely avows, or is disposed to avow sincerely, that p; and
(3) S has at least one want which explains in part both why the belief that not-p is unconscious and why S is disposed to disavow a belief that not-p and to avow p, even when presented with what he sees is evidence against it (p. 137).

So for Audi, the self-deceived individual believes only the true proposition not-P while being disposed to avow the false proposition P because of certain desire(s) she has. This neatly solves the static puzzle since contradictory beliefs are not required – the false belief that not-P is not present.[4] And since Audi appeals to the possibility of unconscious beliefs, he also has a ready answer to the dynamic puzzle in that the true belief that not-P need not enter conscious awareness and interfere with S's ability to "sincerely" avow that P.

Such a view is not very satisfying in light of Sartre's understanding of bad faith, however, nor of the view of self-deception I develop later on. How can we speak of someone as self-*deceived* on Audi's view, if we accept the common-sense view that deception requires a false belief? A person who meets Audi's definition is disposed to "avow sincerely" the false proposition P, but Audi seems to be helping himself to the intuition that we tend only to sincerely avow things that we actually believe. How can we sincerely avow things that we do not believe? That is, an instance of sincere avowal is most straightforwardly an instance of reporting on one's own beliefs, and yet Audi maintains – implausibly – that a belief that P is not necessary for a sincere avowal that P. It is highly tempting to view Audi's view as an *ad hoc* attempt to avoid the force of the static and dynamic puzzles by sidestepping them. Audi admits that he thinks of holding contradictory beliefs as "irrational" and "inexplicable," and so he has

[4] Audi's view prefigures my discussion in Chapter III of the dispositional model of belief. On such a model, beliefs are merely dispositions to behave in various ways. So believing that P disposes the person to, for instance, avow that P when asked. Adopting the dispositional model of belief would allow Audi to hold that self-deception still involves contradictory beliefs, since believing both P and not-P is just to have two different dispositions – and there is nothing obviously paradoxical about that. I endorse just such a strategy to preserve the commonsense model of self-deception while avoiding the static puzzle.

crafted a theory of self-deception that aligns with those intuitions. But it is hardly inexplicable, as Davidson's view demonstrates. Audi may not be respecting a distinction that Davidson (1998) highlights (emphasis mine):

> It is possible for a person to believe contradictory propositions, not only when the contradiction is too subtle for normal detection, but also when the contradiction is obvious (for the contradiction must be obvious if it is to move someone to self deceit). At the same time we should balk at attributing to anyone belief in a plain contradiction. *The distinction we need here is between believing contradictory propositions and believing a contradiction, between believing that p and believing that not-p on the one hand and believing that [p and not-p] on the other* (p. 5).

This distinction, revolving around the content of what the self-deceived individual believes, is crucial to the plausibility of holding contradictory beliefs in self-deception.[5]

Audi's condition (3) also does far too much work in making his theory plausible for how little explanation he gives. If the "want," or desire is so powerful that it keeps the true belief that not-P submerged in unconsciousness and disposes S to avow that P, then how is the desire not powerful enough to bring about a *belief* that P? Once again, the answer must be that Audi does not believe it psychologically plausible. But surely this is an empirical question worthy of an empirically-oriented discussion. In fact, there is copious literature testifying to

[5] Whether holding contradictory beliefs is plausible also depends heavily on what we mean by a contradiction. As Herbert Fingarette (1969) writes in an early book-length treatment of self-deception, "We all no doubt have [contradictory beliefs], if only because we cannot see far enough into the implications of our beliefs...If that were all that were involved in self-deception, who would be puzzled by it?" (p. 15). Fingarette may have in mind commonplace cases like the following: suppose I believe that Superman has no weaknesses while also believing that he seems to sweat copiously around kryptonite. If I were asked whether Superman has a weakness against kryptonite then I could work out, with a little thought, that he has such a weakness. My beliefs prior to reflecting about whether Superman is weak to kryptonite are not "directly" contradictory, but because I can derive a contradiction through further reasoning they are what we might call *indirectly* contradictory. And combined with an empirically plausible model of the mind as divided, a case like this can help explain how direct contradictions arise. If my new belief that Superman is weak to kryptonite arises unconsciously and is hidden away from the belief that he is invulnerable, then we now have a case of genuinely contradictory beliefs (though probably not self-deception, which as Fingarette notes requires more than just a contradiction).

the power of desires to produce beliefs in ways of which we are not aware and lack conscious control – see the next section.

That is why I believe a better alternative to Audi's theory comes from Alfred Mele (1997, 2001). In contrast with Audi, Mele claims that entering into self-deception about a proposition P involves holding the belief that P but need *not* involve holding the belief (or any other kind of attitude) that not-P. But in a similar way, Mele's account is "deflationary" in that his strategy is to avoid the static and dynamic puzzles rather than tackling them head-on. Mele offers the following four conditions as jointly sufficient to describe cases of self-deceptive belief acquisition in a subject S:

1. The belief that P which S acquires is false.
2. S treats data relevant, or at least seemingly relevant, to the truth value of P in a motivationally biased way.
3. This biased treatment is a nondeviant cause of S's acquiring the belief that P.
4. The body of data possessed by S at the time provides greater warrant for not-P than for P.

As an account of the basic phenomenon of self-deception, Mele's theory has much to recommend it. He does not implausibly claim that a false belief is not present in self-deception as Audi does, and so on his theory the false belief can play its important role of motivating the self-deceived individual to protect the deception without requiring that a curiously strong desire do all the work by itself. So of the three accounts we have seen so far, Mele's offers I think the strongest responses to the static and dynamic puzzles; Davidson's and Audi's views, while interesting, have too many problems to endorse as correctly capturing the phenomenon.

If we are keeping score (and I am), Mele's view still does not allow much room for Sartrean bad faith, which I claim is a requirement for a robust theory of self-deception. Mele thinks that self-deception need not be intentional, so it does not fit neatly into the model of other-deception, which bad faith is grounded in. But this is largely because he finds the notion of a divided mind, in which beliefs can be walled off from introspective access, empirically suspect. In Chapter III I argue that there is a plausible empirical story to tell about division, one that I hope Mele would find compelling. And in any case his theory does not *rule out* bad faith as being implicated in self-deception. He simply doesn't think that bad faith is central to self-deception; I hope to muster an argument that would disabuse Mele of that conviction.

Another, related concern is Mele's firm stance on another controversial theoretical question about self-deception: *intentionality*. He holds that self-deception can be "strategic" without being intentional. This is because the "motivationally biased" tactics the self-deceiver uses to deceive herself need not be engaged in "with the intention of deceiving oneself" (p. 98). Mele is rightly concerned that the static puzzle seems to rule out intentional self-deception, since intentional cases must be those in which the self-deceiver's intentions are hidden from him in some sense. "Our ordinary concept of intention leaves room, for example, for 'Freudian' intentions, hidden in some mental partition," Mele writes. He doesn't want to rule out the possibility of such cases (e.g., McLaughlin's "memory stratagem" with forgetful Mary who deceives herself with her diary entries) but wonders why a "paradigmatic" case requires that the subject deceive himself intentionally; he thinks intentional cases are "remote" (p. 100).[6] But, like Audi, Mele underestimates the plausibility of the case for division on the model that Davidson described.

Overall, I consider these worries minor in the context of Mele's great success in giving a concise and elegant theory of self-deception. Mele may lack some imagination about the possibility of intentional self-deception and the plausibility of a divided mind, but I will respond to these worries in later chapters by augmenting Mele's theory, not supplanting it.

[6] Note that Mele thus does end up rejecting a central claim that motivates the dynamic puzzle: he denies that doing something intentionally requires doing it knowingly (i.e., knowing that one is operating under that intention).

Mele's account is also perfectly in tune with my second condition on a good theory of self-deception: empirical sophistication.

1.3 Condition (B): Empirical Sophistication

One major plus of Mele's account compared to others we have seen so far is that Mele is a careful student of the literature in psychology and cognitive science that explains how false beliefs often arise as a result of our desires. According to conditions (2) and (3) of his account (see p. 14 above), "motivationally biased treatment" of evidence is responsible for generating the false belief at the heart of a case of self-deception. Mele deploys the empirical literature in positing that self-deception is not intentional; rather, it arises through various kinds of *biased cognition*, what psychologists call "motivated reasoning." Because motivated reasoning is the process by which false beliefs are generated on Mele's account, I interpret his view as *reducing* self-deception to motivated reasoning: whenever we have a case of self-deception, we have an instance of a false belief produced by unconscious cognitive processes biased by desires; Mele should agree with this interpretation, noting as he does that his theory is "deflationary."

1.3.1 Evidence for Motivated Reasoning

Mele divides biases into two categories: unmotivated and motivated. The latter are what we're really after in discussing motivated reasoning and self-deception, but Mele rightly notes that unmotivated biases can also support self-deception and thus are worth knowing about. Unmotivated biases operate independent of particular desires – they're "always on," you might say.[7] An example is the availability heuristic: beliefs about frequency and likeli-

[7] Mele mentions the so-called "vividness effect" in this category, but literature reviews (e.g., Collins and colleagues' 1986 paper) have found little evidence the effect actually exists.

hood of events are often formed on the basis of how *available* our memory of such events is, and this often fails to reflect the facts. For example, Goldstein and Gigerenzer (2002) discovered that subjects asked to estimate which of a pair of cities was larger in population overwhelmingly chose familiar cities over unfamiliar ones, often correctly.[8] Mele also discusses confirmation biases – we tend to look for evidence that confirms rather than challenges our existing beliefs – and the tendency to search for causal explanations between unrelated events. In a review of the literature, Nickerson (1998) notes that confirmation bias serves as a catch-all for a variety of phenomena. We give "greater weight" to evidence that supports existing beliefs and opinions and often tend only to look for confirming evidence in the first place, he explains, but this can be for any of many different reasons. In some cases expectations and stereotypes do the work, but other cases can be due to things like the *primacy effect*: "People often form an opinion early in the [information acquisition] process and then evaluate subsequently acquired information in a way that is partial to that opinion" (p. 187). Considered as a group these biases are not directed by the agent's desires, Mele argues, since they are operate in a "cold" or unmotivated state.[9] But they nonetheless encourage self-deception by obscuring evidence relevant to the truth or falsity of the belief in question.

Truly *motivated* reasoning occurs when agents misinterpret available evidence to support their pre-existing desires; they selectively gather and emphasize particular pieces of evidence that support P over not-P, even when the evidence for not-P is impartially stronger. Balcetis and Dunning (2006) found that subjects shown ambiguous visual figures "tended to

[8] Biases can often lead to correct conclusions, since they likely evolved as cognitive shortcuts that are "good enough" in many cases.

[9] Mele claims that confirmation bias is truly unmotivated, but Nickerson (1998) distinguishes between motivated and unmotivated varieties. In some cases, yes, we merely seek to confirm pre-existing beliefs in ways that are not influenced by desires. But desires also often play an important role in shaping the way this works. The confusion may lie in how broadly or narrowly we want to understand confirmation bias and which particular experimental results we want to explain.

report seeing the interpretation that assigned them to outcomes they favored" (p. 612). And Dunning (2007) theorized that "self-image motives" like the desire to believe ourselves "capable, lovable, and moral" have significant effects on consumer decision making – "foregoing a medical test or overspending to acquire a set of possessions that is commensurate with one's lofty view of self," for example (p. 247).

For a more vivid illustration, Mele offers up the following thought experiment (1997, p. 94):

> Don just received a rejection notice on a journal submission. He hopes that his article was *wrongly* rejected, and he reads through the comments offered. Don decides that the referees misunderstood a certain crucial but complex point and that their objections consequently do not justify the rejection. However, as it turns out, the referees' criticisms were entirely justified; and when, a few weeks later, Don rereads his paper and the comments in a more impartial frame of mind, it is clear to him that the rejection was warranted.

Even though this example involves reasoning that is motivated by desires, Mele is adamant that it does not show that self-deception is intentional. Rather, our desire to believe P over not-P "primes" bias mechanisms "without our being aware," and those mechanisms then filter evidence that supports our desires (p. 95). Mele thus concludes that "garden-variety self-deception is explicable independently of the assumption that self-deceivers manipulate data with the intention of deceiving themselves, or with the intention of protecting a favored belief" (p. 96).[10] Even if a subject tries to impartially assess the evidence for P and not-P, his psychology sometimes gets in the way. I'll critique Mele's argument about intentionality in Chapter III.

[10] Mele appears to argue that our lack of awareness of the operation of motivated reasoning counts in favor of his claim that self-deception is not intentional. This is not obviously true, though for the sake of simplicity I will not pursue the issue further here.

Motivated reasoning is often innocuous – there are many topics for which a false belief matters little – but sometimes the stakes are much higher. Particularly noteworthy are cases involving beliefs about politics and public policy. In a review of the experimental literature, Glaeser and Sunstein (2013) describe a phenomenon they call *asymmetric Bayesianism*: "the same information can have diametrically opposed effects if those who receive it have opposing antecedent convictions" (p. 1). The phenomenon has significant implications for our understanding of political debates and polarization. One naive view is that presenting accurate, objective information about a controversial topic should help clear up misunderstandings and resolve disagreements. But this isn't what the experiments show at all, as the authors explain: "No less than discussion by like-minded people, such information can led people to have greater confidence and conviction about their antecedent beliefs – and thus to make antecedently divided opinion even more divided than it was before. In short, balanced news can unbalance views" (p. 2). They provide an example:

> People were asked to read several studies arguing both in favor of and against the view that capital punishment has deterrent effects. A key finding was that both supporters and opponents of the death penalty were far more convinced by the studies supporting their own beliefs than by those challenging them. After reading the opposing studies, both sides reported that their beliefs had shifted toward a stronger commitment to what they originally thought. One consequence is that the two sides were more polarized than they were before they began to read (p. 4).

Glaeser and Sunstein survey a variety of findings in this vicinity. Ideological commitments can cause corrections to inaccurate news articles to backfire and produce even greater belief in the inaccurate information, if the article concerns a partisan issue (e.g., Iraq and the existence of an active weapons of mass destruction program). Even on issues with which subjects have no preexisting familiarity, factual information about the issue is interpreted according to political orientation – free marketeers believed that nanotechnology was much less risky than more left-leaning subjects, even though both groups read the same factual information about risks and benefits. In all cases, disparities in convictions about political and social issues guide subjects to make radically different judgments about the credibility of a particular piece of evidence – motivated reasoning at its finest.

1.3.2 Three Insights

This survey of the literature leaves us with a big question: why don't we realize when motivated reasoning is affecting our beliefs, and thus that we are self-deceived? Surely a little effort would get us back on the right track, if we had that kind of introspective access to our thinking. In an important literature review, Ziva Kunda (1990) succinctly explains how exactly motivated reasoning represents a departure from ordinary, unbiased cognition:

> I propose that people motivated to arrive at a particular conclusion attempt to be rational and to construct a justification of their desired conclusion that would persuade a dispassionate observer. They draw the desired conclusion only if they can muster up the evidence necessary to support it. To this end, they search memory for those beliefs and rules that could support the desired conclusion. They may also creatively combine accessed knowledge to construct new beliefs that could logically support the desired conclusion. It is this process of memory search and belief construction that is biased by their goals. The objectivity of this justification construction process is illusory because people do not realize that a process is biased by their goals, that they are accessing only a subset of their relevant knowledge, that they would probably access different beliefs and rules in the presence of different directional goals, and that they might even be capable of justifying opposite conclusions on different occasions (p. 483).

This passage contains three big insights into motivated reasoning. First is that motivated reasoning cannot produce just *any* belief: rather, the set of possible beliefs is constrained by what the available evidence can be twisted to support. This helps explain why self-deceptive beliefs tend to be ones that *would* be justified if the evidence were slightly different. Even if I have a strong desire to believe that E.T. lives deep in my closet and eats the Reese's Pieces I carelessly leave lying around the apartment, absent some truly bizarre background beliefs I probably will *not* be able to deceive myself into believing it. Instead, I will likely come to the reasonable conclusion that my insatiable dog is eating them instead. Alternatively, I could deceive myself into thinking that my gluttonous roommate is eating them, since I really want to preserve my image of the dog as a perfect angel. This false belief, since it can be construed as supported by evidence about my roommate's eating habits, is a much more plausible candidate for self-deception.

Kunda's second insight is that motivated reasoning can involve either mining memory for "beliefs and rules" that the agent already has or constructing new beliefs for the occasion. In either case, though, self-deception seems to require generating a novel false belief. Kunda has identified two ways this can happen. In the first, the self-deceiver already has beliefs and rules relevant to the thing she wants to deceive herself about. Her false belief arises when she wrongly construes those beliefs and rules as applying to the current situation. Returning to the case of the missing Reese's Pieces, I already believe that my roommate is a glutton and so I unconsciously construe that belief as supporting a further belief that the roommate is the culprit. So in this case the false belief is the direct product of another belief I already possess. The second way of generating a false belief via motivated reasoning happens when the agent has no pre-existing beliefs directly relevant to the situation. Suppose I have no roommate and no obvious culprit besides the dog. I might still deceive myself by coming to believe, on the basis of flimsy evidence, that there are rats hiding in the walls who come out and steal the Reese's Pieces away. Assuming I had never before pondered whether rats live in my walls, that false belief would not be the direct product of any preexisting beliefs. These two different ways of generating a false belief should be sufficient to capture the false beliefs at the heart of each and every case of self-deception, I claim.

Finally, Kunda notes the illusion of objectivity that accompanies motivated reasoning. Even though our unconscious is constantly generating beliefs through a biased process that weighs evidence according to our desires rather than standards of impartiality, we have no introspective awareness that this is so. Quite the contrary: we tend to think that our belief-forming processes *do* live up to what impartiality requires, which explains why motivated reasoning – and self-deception – is such a puzzling and frightening phenomenon.[11] Even someone fully informed of the literature on motivated reasoning can only know that certain of her beliefs are the product of bias; she cannot know which ones, at least not with any certainty. So accusing someone of self-deception will necessarily be a fraught prospect, since at first blush the accused will be unable to summon to mind any evidence of biased cognition. In extreme cases, this can make the self-deception incorrigible.

All three of these insights serve to enrich our understanding of Mele's theory of self-deception. Armed with our knowledge of the literature, we are now better equipped to make predictions about the kinds of things people self-deceive about, how those false beliefs are formed, and the extent to which self-deception is invisible to introspection (answer: quite a lot!). By making motivated reasoning the foundation of his theory, Mele demonstrates an admirable, thoroughgoing commitment to empirical evidence in understanding self-deception; this is what I mean when I say that empirical sophistication is a mark of good theory of self-deception.

1.4 Condition (C): Real-World Cases

[11] In another article, Kunda and Sinclair (1999) observe that people under the influence of motivated reasoning "do not wish to be or to appear to be biased by their goals. They are committed to rationality…" (p.13). So motivated reasoning, and self-deception too, are troubling because they involve cognitive processes that run counter to an explicit commitment to rationality, a commitment that many of us tend to hold dear. Motivated reasoning pits one part of our psychology against another.

An important contemporary treatment of self-deception comes from the great evolutionary biologist Robert Trivers, whose 2011 book *The Folly of Fools* describes self-deception as the product of evolutionary adaptation. I don't have the space to critically evaluate this claim here, and it is orthogonal to my larger project in this chapter and dissertation.[12] But Trivers' discussion is eminently valuable because he grounds it in a vast array of real-world cases. The accounts I have surveyed up to this point display a frustrating lack of engagement with the real and serious effects that self-deception has on the world. Self-deception is not just an abstract concern for philosophers to puzzle over but a vivid, pervasive part of real human lives. Philosophers cannot understand and explain self-deception from an armchair.

Trivers dispenses quickly with his definition of self-deception and doesn't dwell on the conceptual issues that bedevil philosophers. He writes that "the key to defining self-deception is that true information is preferentially excluded from consciousness and, if held at all, is held in varying degrees of unconsciousness. If the mind acts quickly enough, no version of the truth need be stored" (p. 9). The crucial bit is understanding what Trivers means by "preferentially excluded." Treading similar ground as Mele, Trivers explains that self-deception is accomplished via the operation of various forms of biased cognition, often unconscious. He classifies nine distinct kinds of self-deception, each arising from different kinds of bias. One is simple self-inflation of personal abilities and attributes, which "results in people routinely

[12] Though his view does have lots of interesting, controversial, and counterintuitive implications. Trivers claims early on, for instance, that a robust form of self-deception likely predates language, raising the possibility that self-deception is a feature of animal lives as well. He explains:

> In nature, two animals square off in a physical conflict. Each is assessing its opponent's self-confidence along with its own – variables expected to predict the outcome some of the time. Those who believe their self-enhancement are probably more likely to get their opponent to back down than those who know they are only posing. Thus, nonverbal self-deception can be selected in aggressive and competitive situations, the better to fool antagonists (p. 13).

To the extent that non-human animals are subject to biases like ours, animal self-deception isn't overly troubling. But as will be come clear in later chapters, paradigmatic self-deception requires sophisticated capacities unique to humans.

putting themselves in the top half of positive distributions and the lower half of negative ones. Of US high school students, 80 percent place themselves in the top half of students in leadership ability. This is not possible" (p. 16). Trivers suggests that effects like these are tied to activation of the medial prefrontal cortex (MPFC), which "seems to be involved in processing self-related information" (p. 17). Lesioning MPFC "delet[es] an individual's tendencies towards self-enhancement," he reports (p. 17).

Another kind of self-deception Trivers calls "the biases of power." He notes that "[w]hen a feeling of power is induced in people, they are less likely to take another's viewpoint and more likely to center their thinking on themselves. The result is a reduced ability to comprehend how others see, think, and feel" (p. 20). Winston Churchill is held up as an example: "At the heights of his power, he was described as dictatorial, arrogant, and intolerant, the stuff of which tyrants are made; at low power, he was seen as introspective and humble" (p. 21).

I'm not sure that these kinds of phenomena are really self-deception, mostly because neither category we've just seen seems to necessarily require a false belief. In the first case, consider that qualities like leadership ability are nebulous – there is no single obviously correct way to go about measuring them.[13] So while the students' positive self-assessments *could* be the product of false beliefs about where they rank relative to their peers, they could also be due to each student's using a notion of leadership ability on which she ranks highly; the decision about how to measure leadership ability might be self-serving, but there could easily be room for reasonable disagreement. My general complaint is that mere bias is insufficient for self-deception, which crucially requires a false belief in addition.[14] This same worry applies to

[13] Trivers is more on-target with an example about self-assessments of physical attractiveness (p. 16).

[14] That's why we can't interpret Mele's theory as saying that motivated reasoning is *identical* with self-deception: there will be many cases of motivated reasoning (and also unmotivated bias) that do not result in a false belief and thus are not self-deception.

Churchill-type cases: even if it's established that having power made Churchill behave in ways very different from how he behaved while powerless, it's highly unclear that Churchill was self-deceived. What was he deceiving himself *about*?

Trivers is on more solid footing with some of his other categories of self-deception. He reviews the compelling literature on the "illusion of control" we have over events around us. Stockbrokers were seated in front of a monitor displaying a "a line moving across it more or less like the stock market average, up and down…, able to press a computer mouse, and told that pressing it 'may' affect the progress of the line, up or down. In fact, the mouse is not connected to anything" (p. 23). Despite the lack of control, many subjects still reported feeling able to influence the line. Why? Trivers hypothesizes that "certainty of risk" is more comfortable than uncertainty, and feeling in control yields certainty (p. 22). Those subjects who reported a high illusion of control tended to be judged less valuable employees by their firms and were paid lower salaries. Trivers offers that this is perhaps because knowledge of the limitations of one's control is part of being a successful stockbroker. Whatever the explanation, a tendency towards self-deception creates the potential for significant negative consequences.[15]

One other rich category of self-deception revolves around what Trivers terms "biased social theory:"

> Are the wealthy unfairly increasing their share of resources at the expense of the rest of us (as has surely been happening) or are the wealthy living under an onerous system of taxation and regulation? Does democracy permit us to reassert our power at regular intervals or is it largely a sham exercise controlled by wealthy interests?…And so on (p. 24).

[15] This is not to say that self-deception is always or even often is a bad thing. The account of self-deception I will develop is value-neutral, and we will see various cases where self-deception seems to offer significant benefits.

How we answer questions like these depends largely on how we make sense of a complex web of facts – how we sift through and ultimately interpret the available evidence. To the extent that our interpretations are false beliefs driven by bias, the social theories through which we view the world are grounded in self-deception.

The latter part of Trivers' book is devoted to examinations of self-deception in important social and historical contexts. He spends an entire chapter illuminating the issue of false historical narratives, "lies we tell one another about our past" for reasons of "self-glorification" and "self-justification" (p. 215). The tale of Christopher Columbus hits particularly close to home for Americans. Columbus is revered in U.S. history as an explorer who discovered the "New World" and began the process of taming the land that would become America; he even has a holiday named for him! But his treatment of the native Indians was horrific: "newborns given to dogs as food or smashed against rocks in front of their screaming mothers, twenty thousand killed in Hispaniola alone, with more to come on nearby islands" (p. 219). We prefer to forget that things like this happened, and in remembering the past this way "we deny the motives and the reality of territorial takeover," Trivers soberly assesses (p. 219).[16]

Elsewhere, Trivers emphasizes the role of self-deception in various aviation and space disasters. The 1986 *Challenger* explosion, according to an analysis by the physicist Richard Feynman, was partly caused by self-deception within NASA. Having already beaten the Russians to the moon, the $5 billion space program needed to justify itself, and so "critical design features, such as manned flight versus unmanned flight. were chosen precisely because they were costly" (p. 201). When 7 of 23 earlier *Challenger* flights showed damage to safety-critical

[16] It's true that many Americans may not know the real facts about Columbus well enough to count as engaging in self-deception themselves. But Trivers' point is that those who *created* the false, self-serving narrative about Columbus almost certainly were. And in the aggregate, argues Trivers, "false historical narratives act as self-deception at the group level" (p. 246).

O-rings during low-temperature takeoffs, NASA's safety unit disregarded the other 16 flights (which had higher-temperature takeoffs) as irrelevant to determining the cause of the damage – a flagrant violation of good statistical practice. This caused the engineers to miss the fact that low temperatures were associated with higher damage, since they did not compare the 7 high-damage, low-temperature flights with the no-damage, high-temperature flights. Trivers argues that, under pressure to keep the program running, NASA engineers deceived themselves about how to analyze the *Challenger*'s O-ring safety issues. Since the explosion occurred during a low-temperature flight, their self-deception is directly implicated.[17]

Finally, and not uncontroversially, Trivers discusses how certain religious practices may also be grounded in self-deception. Consider the Christian belief in the power of intercessory prayer, which Trivers cheekily describes as "people in a room, scrunching up their foreheads in intense concentration on behalf of someone miles away about to undergo surgery" (p. 299). A large, well-controlled study involving six hospitals concluded that intercessory prayer has no effect on health outcomes. And, ironically, patients who knew they were being prayed for had more post-operative complications than patients who did not. Yet belief in prayer's effectiveness endures. How to explain this, other than by self-deception?

A hazard of using real-life cases like these is that there is often room for reasonable disagreement on the interpretation – perhaps the *Challenger* safety issues were the result of a genuine mistake, for instance. But the sheer quantity and variety of such cases tells against explaining away *all* of them. At least some of the cases Trivers discusses surely get right at the heart of the phenomenon I am after in this dissertation: genuine instances of self-deception that are large-scale and pervasive in ways that lay bare the profound influence self-deception

[17] This may not be exactly the way Trivers interprets the case. Instead he may believe that NASA itself was self-deceived and the engineers helped facilitate the whole organization's self-deception. I think the engineers themselves were plausibly self-deceived, and this fits better with the model I develop in the next chapter.

can have on the things we care most about. Indeed, things that are important to us are even more likely to be the topics of self-deception, given the strength of the desires that surround them – when a belief matters dearly to me, I tend to be highly invested in its being true.

So the crux of condition (C) is that self-deception is fundamentally a natural phenomenon first: investigation of self-deception should proceed by gathering together large classes of similar-seeming, real-world cases and then finding a philosophical explanation that unifies them. Many philosophers, I think, tend to approach the subject backwards, beginning by weaving intricate theoretical models and only later bringing in examples and thought experiments. The risk with that approach is that you can end up with an elaborate theoretical apparatus that is only tenuously connected to anything going on in the world around you that you might want to *actually* explain. Mele is not wholly unsuccessful on this count, since his theory is grounded in the empirical literature on motivated reasoning; he clearly believes that traditional philosophical approaches to self-deception are insufficient. But he does not go far enough towards explaining why self-deception *matters*. I argue that self-deception is a big deal – more than Mele realizes – because of the frequency and magnitude of the self-deception that exists around us.

1.5 Looking Ahead

I have argued for three conditions that I believe should be a captured by a full theory of self-deception like the one that I offer in this dissertation. They are: (A) preservation of Sartre's "bad faith" intuition, (B) empirical sophistication, and (C) engagement with real-world examples. As the conclusion of a survey of several prominent views of self-deception, I have also claimed that Alfred Mele's account provides the best general theory in the extant literature. It satisfies (B) and with the revisions and augmentations I offer in the next chapters will satisfy (A) and (C) also.

This has been the first of four chapters. In Chapter II I describe the general characteristics of "paradigmatic" cases of self-deception and explain both why they are important and how Mele's theory of self-deception cannot give us a satisfying explanation of them. Chapter III argues that paradigmatic cases involve *intentional* self deception, which Mele and others claim is rare or even impossible; in particular, I claim that paradigmatically self-deceived individuals who falsely believe that P are intentionally ignorant of the true belief not-P (which I claim they also possess by virtue of a "partitioned" mind) via intentional manipulation of their attentional capacities. Finally, Chapter IV uses intentionality to argue that in paradigmatic self-deception the self-deceiver is apt for judgments of moral responsibility, which extant theories like Mele's leave little room for. This completes the portrait of paradigmatic cases as a distinctive, morally interesting class that has heretofore gone unexplored in the philosophical literature on self-deception.

Chapter II
The Paradigm of Self-Deception

In the days leading up to the 2012 U.S. presidential election, polls were mostly unanimous in giving Democratic nominee Barack Obama a slight but definite lead over his Republican opponent, Mitt Romney. Statistician Nate Silver, writing for the *New York Times* on election eve, gave Obama a 90.9% chance of winning based on an aggregate analysis of the latest polling numbers; an Obama victory was exceedingly likely (Voorhees 2012).[1] And so the results followed the polling closely: Obama eked out victories even in heavily contested swing states like Ohio and Florida, and NBC News called the race at 11:12 p.m. ET – slightly later than in 2008 but much earlier than in 2000 or 2004.

Yet despite the polling evidence, the Romney campaign was seemingly blindsided in its defeat, having believed that the polling was simply wrong. As CBS News reported, "[The campaign] believed the public/media polls were skewed – they thought those polls oversampled Democrats and didn't reflect Republican enthusiasm. They based their own internal polls on turnout levels more favorable to Romney" (Crawford 2012). The Boston Globe reported that the campaign spent $25,000 on a planned victory fireworks celebration (Johnson 2012). Romney, exuding confidence, told reporters the day before the election that he had prepared a victory speech but not a concession statement (Weiner 2012). Lest we think this was mere posturing, CBS News reported that Romney's concession remarks on election night were

[1] Silver rose to fame as an expert in sabermetrics – statistical analysis of baseball – before venturing into politics with his blog FiveThirtyEight in the 2008 presidential election, for which he correctly predicted the electoral results in 49 of 50 states. Besides Silver, other statisticians were also vindicated by the 2012 results: David Rothschild of the blog PredictWise correctly predicted (Rothschild 2012) a 303-vote Electoral College total for Obama back in February of 2012 – some nine months prior – and Sam Wang of the Princeton Election Consortium called 49 of 50 states correctly (Wang 2012).

“hastily composed” just moments before; it wasn’t a joke (Crawford 2012). That same CBS News report provided an eye-opening account of the campaign’s mindset in its final hours (Crawford 2012):

> "We went into the evening confident we had a good path to victory," said one senior adviser. "I don't think there was one person who saw this coming." They just couldn’t believe that they had been so wrong.

It was perhaps even worse for many conservative Internet bloggers, who devoted considerable energy to “debunking” the mainstream polls as hopelessly biased in favor of the Democrats. One, a man named Dean Chambers, authored a highly trafficked site called UnskewedPolls.com, where he reweighted polls from organizations like Gallup and the major television networks, which were unfavorable to Romney, to match the weighting of Rasmussen Reports, a conservative polling group that was virtually alone in forecasting strong Republican turnout and a likely Romney victory. In the end, the mainstream polls were right and Chambers was wrong. While a great many people were self-deceived about the polling, for the sake of simplicity I’ll focus just on Romney and his campaign going forward.

2.1 What Went Wrong With Romney?

What is interesting with cases like these is not that the Romney campaign held false beliefs: many people are wrong about a lot of things a lot of the time, you might think, and that is not overly concerning. Instead, what is most interesting is that they were *confident* in those false beliefs right up to the end despite contrary evidence that was, to a disinterested observer, vividly available the whole time. Surely the deluge of contrary evidence warranted some caution: a reasonable person, you might think, would have been much more careful. Yet caution and care were nowhere to be seen, and consequently, there also seems to be room for *blame*. Mitt Romney and his campaign made serious mistakes – ones we can hold them morally responsible for. Put another way, those donors who paid for Romney's wasted fireworks would be justifiably angry. The same goes for anyone who invested money based on the Romney's promises to lower taxes upon election or bet against Obama in online markets like Intrade, and for those who spent countless hours rebutting Obama supporters on the basis of Chambers' faulty assumptions.[2] Romney misled many people as a direct result of his false beliefs, and this tends to make people upset. But are those feelings of anger and resentment *justified*? The question turns on whether Romney is properly morally responsible.

2.2 Prelude: Responsibility

I will approach moral responsibility from within the framework of P. F. Strawson's landmark 1962 paper "Freedom and Resentment," in which Strawson argues that holding others responsible is a matter of expressing certain "reactive attitudes," sentiments like anger, gratitude, and forgiveness. In contrast to other theories of responsibility, Strawson's account finds the justification for holding others responsible not in any substantive metaphysical commitments but in the *role* that the reactive attitudes play in society. We are all embedded in various interpersonal relationships – bonds of friendship, family, society, and so forth – and in the course of those relationships people act so as to exhibit attitudes of "good will, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other" (p. 5). We cannot help but respond to those actions in certain ways, says Strawson, namely by expressing reactive attitudes which themselves constitute moral judgments.

By itself this account is underdeveloped and wide open to objections. Various authors, including Daniel Dennett (1984), R. Jay Wallace (1996), and more recently Manuel Vargas

[2] Some responsibility surely falls on Romney's and Chambers' supporters for being too credulous – they may have been self-deceived too – but Romney and Chambers bear causal and, I claim, moral responsibility as leaders of their respective groups.

(2004), have tried to defend it against serious charges including what Vargas calls the *cognitivist criticism*: "This is the idea that claims of responsibility admit of truth and falsity, and our 'grounding beliefs' – the beliefs that provide the foundations for our judgments that someone is responsible – might well be false" (pp. 221-2). That is, responsibility judgments can be true or false because they rely on beliefs that can be true or false; we might use reactive attitudes in ways that are mistaken about the metaphysics of responsibility.

This debate remains unresolved, and arguing about responsibility remains notoriously difficult to do authoritatively, but for the purposes of my argument it suffices to note that a robust Strawsonian tradition holds that we can analyze responsibility by discussing reactive attitudes; my argument going forward will operate within that tradition.

It may also help to emphasize that from one perspective holding others responsible is not a theoretical matter but a practical one. To return to Romney and his campaign, we are embedded in a certain form of political system here in America, and there is a practice of holding responsible those who make certain kinds of mistakes within that system. We can and do disagree about what metaphysical commitments underlie these practices – the broad dilemma between Strawsonian and more substantive theories of responsibility seems to have reached a stalemate – but it is I hope obvious that Romney's behavior at least raises the *issue* of responsibility. He and his campaign made catastrophic mistakes and misled themselves and others in a way that none of us would want to emulate; many people can and would blame them.

2.3 The Notion of a Paradigm

Given a broadly Strawsonian view of responsibility, you might wonder why the Romney campaign's mistakes are uniquely *apt* for responsibility judgments. After all, we all hold false beliefs – hundreds of years ago everyone believed that the sun revolved around the earth – and yet we only hold each other responsible for some of them. Which ones? In Chapter I, I discussed the importance of motivated reasoning for the phenomenon of self-deception. Our tendency to automatically and unconsciously interpret evidence in biased ways – those motivated by our desires – leads us to form false beliefs that are the basis of self-deception.

I also presented Alfred Mele's influential "deflationary" view of self-deception, which I read as claiming that synchronic instances of motivated reasoning are sufficient to count as instances of self-deception. This sort of view poorly explains what went wrong with people like Romney, or so I claim. Deflationary views fail to provide insight into what about such people's behavior was so egregious and so also fail to motivate the aptness of responsibility judgments. I need an alternative.

In this chapter I will develop the intuition that Romney's case is part of general class of cases that I call *paradigmatic cases*. These cases are instances of self-deception that I claim are most central to the phenomenon and its understanding. I define paradigmatic cases as those displaying three interrelated features: (1) the deception is *diachronic*, (2) the relevant false belief deviates from reality to a high *degree*, and (3) sustaining the deception requires *effort*. All three features come in degrees, and none are strictly necessary or sufficient for a case to fit the paradigm of self-deception. ("Fitting the paradigm" here means looking at a case and judging the degree to which it displays the three features, which I think collectively characterize central or "paradigm" cases of self-deception.) Also note that while each feature plays a valuable epistemic role in identifying instances of self-deception – that a case of self-deception is diachronic, deviates greatly from reality, and is effortful lets us more easily identify it as such – I am here claiming that each feature is also *characteristic* of the paradigm of self-deception.[3]

Let's now examine each feature in turn. My strategy will be to consider how the Mitt Romney campaign displayed each of the three features and to argue that deflationary theories are ill-equipped to explain how each contributes to the severity of the self-deception. With the notion of a paradigm in hand, chapters IV and V will trace the route from paradigmatic cases to moral responsibility.

2.4 Diachronicity

As I argued in Chapter II, motivated reasoning often leads us astray when we try to evaluate evidence for and against our beliefs. On accounts like Alfred Mele's, false beliefs produced this way are sufficient to produce self-deception all by themselves. This seems correct from one perspective: my own psychology conspires against me – though in ways of which I'm not consciously aware – to produce a deceptive, false belief. Yet while this definitely occurred in the case of Romney, it is far from the whole story.

For one thing, the self-deception was for them an ongoing *process*; its temporal extension seems key to explaining why the self-deception is so egregious. Suppose that on July 27, 2012, candidate Mitt Cromney and his team receive some potentially bad news. The critical swing state of Ohio does not appear to be headed in a favorable direction; a new poll from Magellan Strategies shows Obama ahead by two points, potentially indicating that Ohio's sharp drop in unemployment is paying dividends for the President (Silver 2012). Moreover, the poll is merely the latest in a long line of similar results – the campaign's efforts don't seem

[3] An analogy is Richard Boyd's (1988) notion of *homeostatic property clusters*: the paradigm of self-deception is composed of such a cluster.

to be moving the needle.[4] But the Cromney campaign has commissioned its own poll of Cromney that shows the opposite: Cromney is doing well and at present has a good chance of winning the state. However, the campaign's in-house pollsters constructed the poll using a voter turnout model especially favorable to Cromney – an assumption that is at best controversial and at worst fatally optimistic. A senior campaign adviser presents the two polls to Cromney, assuring him the campaign's poll is accurate while the Magellan poll is biased towards Democrats. Desiring to believe that he is doing well in Ohio, Cromney then forms, as a result of motivated reasoning, the false belief that his campaign's poll is more accurate than the Magellan poll. On Mele's and other accounts, this suffices for self-deception. Suppose, though, that after the adviser's presentation a junior Cromney staffer who works on polling research speaks up with another perspective. He offers an impromptu report arguing that the campaign needs to temper its optimism and revise Republican turnout projections downward. Cromney then takes the report seriously and rejects his earlier belief that the campaign's polling is solid. So his false belief here ends up being short-lived: motivated reasoning led Cromney astray, but he soon corrected the error by seeking out more information.

The most interesting, vivid cases of self-deception seem to be much more temporally extended than the one we've just seen. Suppose that in another world – one more like the actual one – candidate Mitt Romney sticks with the campaign's problematic polling assumptions and continues to disbelieve unfavorable polls from sources outside the campaign. Over the many months before the election, he continually evaluates outside polling in a biased way that protects his favored belief that the other polling is based on wrong assumptions; any naysay-

[4] This trend – a small but steady Obama lead – continued all the way up to the election, in Ohio and most of the crucial swing states. While it is always possible for one poll to be mistaken or even for the polling in a whole state to be badly calibrated, it was extremely unlikely that the swing state polling was off across the board *and* in a way that was unfavorable to Romney. Yet in constructing its polls the campaign seemed to make just this assumption.

ers on the staff are made to feel unwelcome. On election night he is blindsided by his loss, never having taken seriously any evidence that could have challenged his false belief.

This too is self-deception. But surely these two cases are highly dissimilar: in the one-poll case there is an instance of simple motivated reasoning soon corrected, while in the many-polls case Romney is engaged in an ongoing process of mis-evaluating evidence to preserve a false belief – a belief that is never corrected. A Melean account of self-deception captures some of the intuitions important for understanding self-deception: having a false belief is certainly a critical part of being self-deceived, and motivated reasoning is important for explaining how such beliefs arise. But his account does not give us the resources to say why Romney's months-long disregard for evidence about the campaign's polling problems, on the one hand, is more severe than Cromney's transient misappraisal of the evidence on the same issue.[5]

Instead of saying that both are simply instances of self-deception, as Mele would, I think we need a theory of self-deception that gives us the resources to explain why Romney‘s case is so different from Cromney's. Cromney and Romney's cases begin in precisely the same way, but Romney's false belief about the reliability of his campaign's polling has persistent and vivid effects on his behavior; his false belief in this case is the genesis of a diachronic pattern of distorting relevant evidence. Because of this distinction, I argue that that the case

[5] It is important to emphasize that I do not think that self-deception must manifest in outward behavior. The beliefs relevant to the self-deception may manifest internally (i.e., in the subject's psychology) as well as – or even instead of – externally, and such cases are just as vividly instances of self-deception as cases in which the self-deception is obvious to a casual observer. My general point is merely that paradigmatic cases of self-deception are diachronic. This is not true of Romney.

of Romney belongs in a different category from those that Mele has in mind.[6][7] On my view a theory of self-deception that lumps them together is both misleading and inaccurate; self-deception is more fine-grained than Mele thinks because there *are* important differences between the two cases, and so the notion of a paradigm of self-deception that allows us to draw distinctions between them is sorely needed.

2.5 Degrees of Deception

Another factor that points to agency in paradigm cases of self-deception is the degree to which the agent's false belief deviates from reality. Suppose now that Mitt Cromney meets with his campaign staff to discuss their polling and an adviser presents him with two options. On the one hand, they can use a turnout model favorable to the candidate: crowds for Cromney's speeches have been large and enthusiastic in swing states while Obama has struggled, suggesting that Cromney's message is connecting with many crucial voters. But the alternative is to use a model that matches the one from 2008: even though many voters are unhappy with Obama's management of the economy, they still find him likable and trust that he's doing the best he can – Cromney's base is more enthusiastic than Obama's but this won't necessarily translate into greater turnout. The best available data shows that the favorable model is 49.9% likely while the unfavorable is at 50.1%, so Cromney faces a tough decision. He believes that respecting the facts is important and knows that political campaigns often go disastrously wrong when polling is divorced from reality. But Cromney has also long been an optimist in

[6] Even a false belief lasting a few seconds would be diachronic in a trivial sense, of course, so I acknowledge that there is no (non-arbitrary) sharp line demarcating cases of paradigmatic self-deception from ordinary cases of having a false belief under Mele's definition. After all, if Cromney in the space of an afternoon considers and rejects the bad turnout model, then it seems quite odd to call that self-deception at all; accusing someone of self-deception seems different from accusing her of motivated reasoning.

[7] Also note that my claim is *not* the trivial observation that it's easier to identify cases of self-deception by looking at behavior across time. Rather, I claim that paradigmatic self-deception is *itself* diachronic.

all areas of his life. He took many risks in his personal and professional life on the road to being the Republican nominee, and many of those risks were taken even when the odds were against him; this has earned him the respect and admiration of many. In the end, Cromney the optimist takes one more risk and chooses the favorable turnout model. That model turns out to be wrong, and on a Melean account Cromney is self-deceived since the evidence objectively weighs against his belief.

Through the looking glass, Mitt Romney faces a similar choice except that polling experts tell him that the favorable model is only 30% likely while the unfavorable model sits at 70%. Romney goes with his gut and stuns his advisers by picking the favorable model. He goes on to lose, to the surprise of no one but Romney himself. This case again involves a false belief acquired in opposition to the evidence at hand; it too is self-deception on a Melean theory.

How to tell these two cases apart? Romney is clearly self-deceived, but Cromney seems to be engaged in something else: *wishful thinking*. The distinction between self-deception and wishful thinking is controversial. Mele (1997) thinks of wishful thinking as on a continuum with self-deception: "[T]he difference may lie in the relative strength of relevant evidence against the believed proposition" (p. 100). Others, like Dion Scott-Kakures (1996), argue that self-deception and wishful thinking are distinct phenomena because only the former involves an intentional action. I don't find either of those explanations satisfying. Scott-Kakures is right that wishful thinking seems a different phenomenon, but it's not clear that intention is what makes the difference.[8]

[8] Especially not the intentions that Scott-Kakures has in mind. As I discuss later on, he views self-deception as a special kind of intentional irrationality. I think this approach is wrongheaded, so it is not a sound basis for demarcating self-deception from wishful thinking. In the next chapter I explore how self-deception could be voluntary and perhaps intentional in some sense.

Like Mele, I think that Cromney and Romney fall along a spectrum in that the evidence against their beliefs differs in degree but not in kind and in that sense wishful thinking – assuming that's what's going on with Cromney – and self-deception are of a piece. But a spectrum model doesn't preclude the notion that the poles of the spectrum represent distinct categories; there are borderline cases in the middle, but towards the poles we can readily distinguish wishful thinking from self-deception.[9] Cromney is an optimist who is slightly more confident in the favorable turnout model than the evidence warrants. Romney displays a flagrant disregard for the overwhelming evidence that the favorable model is the wrong choice: at best, he's *pathologically* optimistic.

We have thus arrived at a distinction between wishful thinking and self-deception – or *paradigmatic* self-deception, at least. I'll grant to Mele that the two are alike in many respects and that attempts to draw a sharp line between them are futile, but we can distinguish them to the extent that we can distinguish between, in this case, normal and pathological optimism. Paradigmatic self-deception is revealed when a false belief's deviation from reality reaches extreme levels.

2.6 Effort

Lastly, self-deception seems to be *hard*. Romney and his staff didn't stumble into self-deception by accident. The evidence against their polling model was real, and lots of people saw it and responded appropriately. The effort required for paradigmatic self-deception seems doubly high when you look at the two features we've already seen: when a motivated false belief is sustained over a long period and the belief is stridently at odds with the evidence, then some serious effort is required to preserve that belief.

[9] I have in mind here pairs like hot-cold, tall-short, young-old, etc.

What does this effort look like? I claim that paradigmatic self-deception requires effort both to maintain focus on the evidence that (purportedly) *supports* the false belief and also to inculcate habits of avoiding evidence that *contradicts* the belief. Romney elected to trust his own campaign's turnout model over third-party alternatives and was self-deceived in doing so. Suppose a young staffer speaks out against this decision in a September meeting. Election day is drawing closer, and mainstream polling results aren't moving towards the campaign's own. "Our bad polling is ruining our chance to beat Obama! Sir, I beg you to reconsider!" he exclaims with outrage bringing up the latest unfavorable Gallup poll of Ohio to Romney. But Romney waves the poll away and launches into talking about how much enthusiasm he's seen at recent Ohio rallies and about the many stories he's heard from ordinary Ohioans about how frustrated they are with the president and his policies. "So you see," he concludes with a smile, "there's no reason to worry: the polls aren't showing what we're seeing on the ground." The staffer, embarrassed, nods silently and the meeting continues.

In this situation Romney applied effort towards calling to mind all the evidence that makes his campaign's turnout model seem like the right one. He took a break from the normal course of the meeting to give a speech about why the dissenting view was wrong, and he did so in a way that reminded himself and everyone present of the rationale for the belief that was guiding their polling. Shutting down the objection quickly and completely was crucial to maintaining the campaign's conviction that it was on track for victory.

Obviously this has a practical dimension: doubts among the staff about core strategy decisions make them less effective employees. But the speech serves just as much as a tool for Romney himself to sustain his false belief. Without such tools, the self-deception would threaten to collapse. That explains why Romney would take time out from a busy meeting to repeat things that everyone present already knows about – it's the *reminder* that's important.

So one facet of effort in self-deception is deploying tools that allow the agent to maintain focus on the evidence supporting the relevant false belief through reminding her of the evidence's importance.

Focusing on evidence that supports the belief is only half the story. What about threats from evidence that directly contradicts the belief? Imagine that our contrarian staffer is only briefly chastened by Romney's dismissive attitude – he just *knows* he's on to something, even if no one else in the campaign sees it. When he can, he chats with members of congress, advisers, and friends of Romney and tries to convince them that the campaign is making a terrible mistake. He finds some of them receptive and urges them to call Romney directly. This catches Romney off guard: people he thinks are calling to offer support are instead offering increasingly dire warnings and threatening to tell their friends that the campaign has its head in the sand. Romney is annoyed and also distracted; he wants to keep his attention focused on giving energetic speeches and bringing in big donations, and this kind of advice isn't helping. To avoid such unpleasantness Romney asks his overburdened personal assistant to begin screening all his calls more strictly. The assistant and the callers are unhappy about it and this creates even more stress for Romney, but it's worth it to him to eliminate such a big distraction.

Meanwhile, the staffer continues to speak up in meetings whenever the subject of polling comes up, doing research ahead of time on the latest polls to have facts at the ready. Romney grows impatient: something must be done. The staffer's contributions to other parts of the campaign are valuable; he's too important to fire, Romney decides. Instead, Romney decides to exclude the staffer from most meetings and have him submit written reports on his work for the campaign. This is more work for Romney: he or his assistant has to read all these reports in the vanishingly little free time Romney has – time he'd rather spend with his family.

But this decision solves the problem and soon Romney again feels even surer of the campaign's direction.

Tamar Gendler (2007) calls avoidance strategies in self-deception such as these as ways to avoid *evidential override* – a situation where the evidence against the self-deception is so vivid as to "override" the deception and force it to collapse. As she puts it, "[T]he self-deceived subject deliberately avoids putting herself in situations where she suspects she may come upon evidence wildly incompatible with [her false belief]" (p. 244). This deliberate avoidance requires effort. Even if the effort involved is relatively minor – suppose Romney was able to fire the contrarian staffer with no ill effects – the actions taken still represent time, energy, and other resources expended *solely* in the service of sustaining the false belief. When the threat from contrary evidence is continual, as in the above cases, the effort required to avoid it is greater.

2.7 Agency and Reflection

I have now argued that paradigmatic self-deception is associated with effort both to sustain focus on evidence favoring the relevant false belief and to avoid evidence that contradicts it. Together with the other two features we now have a better idea of the general characteristics shared by paradigmatic cases. Paradigmatic self-deception turns out to be an effortful, temporally extended *activity* involving a false belief that is sharply at odds with reality. How do we move from this characterization to the conclusion that moral responsibility applies? One natural suggestion is that moral *agency* is involved: paradigmatic self-deception is an agential activity, the kind of thing we do as moral agents. Since responsibility is often thought to somehow track instances of agency, establishing that paradigmatic self-deception is agential would make establishing the aptness of responsibility judgments reasonably straightforward.

As it turns out, not many philosophers have argued for an agential theory of self-deception. An important exception is Dion Scott-Kakures (2002), who claims that self-deception is an agential activity and that deflationists like Mele fail to respect the distinction between self-deception and mere motivated reasoning. Scott-Kakures motivates his account by asking us to consider a cat called Bonnie (pp. 578-9):

> Like most felines, Bonnie can make fine aural discriminations. She can, for example, distinguish the sounds of the removal of her own medication from the cupboard from the sounds of the removal of other objects – she promptly disappears only when her own medication is removed. Bonnie is also exceedingly fond of her food. She is apt to scamper into the kitchen when she hears a can of her food being opened. She rarely scampers into the kitchen when some non-cat-food item is being opened. Yet, on occasions upon which Bonnie is *very* hungry she does certainly appear to mistake non-cat-food sounds for cat-food sounds; she scampers into the kitchen when, for example, the coffee grinder is removed from the cupboard. On occasions when she is not so very hungry she does not mistake such sounds for cat food sounds. I assume that Bonnie enjoys some form of beliefs and desires, and so I can see no great harm in describing the situation so: when Bonnie is under the influence of a strong desire for food, she is apt to come to falsely believe that cat food is currently being opened in the kitchen.

On a Melean account of self-deception, Bonnie is self-deceived about whether her food is being opened in the kitchen. Scott-Kakures wants to resist that characterization: surely there is something deep about the difference between the psychological capacities of felines and humans. A good account of self-deception, Scott-Kakures thinks, should be able to explain how self-deception requires a more complex psychology.

Scott-Kakures' first move is to claim that self-deception requires *conceptual sophistication*: "[I]t is by virtue of her lack of such complexity and the associated abilities that the process whereby Bonnie's belief is installed cannot count as self-deception" (p. 583). What does this sophistication amount to? Scott-Kakures paints self-deception as involving a certain kind of error of self-knowledge, one in which the agent "*fails* to make a high enough estimate of the causal role played by [...] desire in the generation of belief" (p. 584). Bonnie lacks the ability to form any judgments at all about the quality of her belief-generating mechanisms – it's unlikely that she has any second-order beliefs whatsoever – so she cannot be said to have failed to understand her own cognition in the way that Scott-Kakures thinks is distinctive of self-deception. Put another way, Bonnie is not a "reflective critical reasoner" who can "be moved by reason *qua* reason, by the thought that *this* is a reason for believing *that*" (p. 585).

So Bonnie cannot be self-deceived because she lacks certain distinctively human reflective capacities. But why think those capacities are important for self-deception? Scott-Kakures claims that in paradigm cases of self-deception the agent "must repeatedly search for evidence and probe various hypotheses in order to arrive at the favored belief and to sustain that belief against threat once it has been generated" (p. 588). She "self-consciously reasons" to a false belief and "evaluates the probity of evidence – she plays an active role in bringing about her error" (p. 590). Those activities are supposed to require reflection, which for Scott-Kakures is precisely what agency consists in: "[T]he stages or steps whereby [the agent] arrives at her belief are mediated by reflective reasoning" because she is "self-consciously seeking doxastic closure with respect to a question of the form 'p or ~p'" (pp. 589-90). That is, agency is part of paradigmatic self-deception because reflective, self-conscious deliberation is an agential activity and that activity underlies the formation and maintenance of the false belief.

2.8 Skepticism About Reflective Agency

I believe that Scott-Kakures' grounding self-deception in reflection is a serious mistake. One reason is that Scott-Kakures seems to be helping himself to the troubling assumption that agency and reflective capacities go hand in hand. This tradition stretches back at least to Kant, who thought that rationality – which operates via reflective deliberation – distinguishes us from lesser creatures and forms the basis of morality. Here is a notable contemporary Kantian, Christine Korsgaard (2009, p. 19):

> [H]uman beings differ from the other animals in an important way. We are self-conscious in a particular way: we are conscious of the grounds on which we act, and therefore are in control of them. When you are aware that you are tempted, say, to do a certain action because you are experiencing a certain desire, you can step back from that connection and reflect on it. You can ask whether you should do that action because of that desire, or because of the features that make it desirable. And if you decide that you should not, then you can refrain.

The empirical threat to this sort of view of agency comes from the psychological literature on *automaticity*, or the theory that human behavior is often – even mostly – determined by automatic, unconscious processes and not controlled, conscious ones like reflection. Psychologists like John Bargh (2005) and Daniel Wegner (2002, 2005) have been at the forefront of a movement to reject the traditional view that conscious willing is causally and explanatorily implicated in our actions, and their work has been sufficiently influential for psychologists like John Kihlstrom (2008) to decry it as creating "the automaticity juggernaut" that has taken over the discipline. Significant controversy exists over how best to interpret the experiments that automaticity proponents cite: Jack and Robbins (2004) accuse Wegner both of being wrong and of actually impeding "the progress of psychology" by requiring us to dismiss our subjective experiences. Bargh's research on social priming effects, which is critical to his theory, has become embroiled in controversy over repeated failures to replicate his findings (Harris *et al.* 2013, Doyen *et al.* 2012). Worries about automaticity may turn out to be overblown, in that conscious willing may play a much larger causal role in our actions than Bargh and Wegner will admit.

Yet even if conscious willing – reflection – is widespread and causally efficacious, critics still readily admit that the *feeling* we have of willing our actions doesn't necessarily correspond to the actual cause. As Jack and Robbins (2004) note, "[Wegner] is also right to identify a folk-psychological belief in direct access as embedded in Western culture and as exerting an influence on Western political, ethical, and legal thought" (p. 665). Eric Schwitzgebel's marvelous article "The Unreliability of Naive Introspection" (2008) explores the general difficulty of drawing definitive conclusions about our phenomenology (p. 247):

> Most people are poor introspectors of their own ongoing conscious experience. We fail not just in assessing the *causes* of our mental states or the processes underwriting them; and not just in our judgments about nonphenomenal mental states like traits, motives, and skills; and not only when we are distracted, or passionate, or inattentive, or self-deceived, or pathologically deluded, or when we're reflecting about minor matters, or about the past, or only for a moment, or where fine discrimination is required. We are both ignorant and prone to error. There are major lacunae in our self-knowledge that are not easily filled in, and we make gross, enduring mistakes about even the most basic features of our currently ongoing conscious experience (or "phenomenology"), even in favorable circumstances of careful reflection, with distressing regularity. We either err or stand perplexed, depending–rather superficially, I suspect–on our mood and caution.

And so, to the extent that reflection and reflective deliberation require the use of a potentially unreliable process, introspection, we should be wary of placing too much trust in them in explaining self-deception. More importantly for my argument in this chapter, we should worry about whether reflection is really a good basis for a theory of paradigmatic self-deception. If there is any substance at all to skeptical worries about reflection, then a reflective theory of self-deception becomes an extremely costly view to endorse.

Suppose we think that even this kind for skepticism about reflective agency is unwarranted. Even then, I think we should resist the move to implicate reflection in self-deception. This is not least because reflection is unnecessary for understanding the phenomenon. I have claimed that paradigmatic self-deception is associated with three distinct features: diachronicity, degree, and effort. We can explain – and I have explained – all three features without invoking reflection or reflective capacities. Diachronicity and effort are strictly behavioral: the agent sustains a false belief over a long period of time and in an effortful way.[10] Degree is a condition on the false belief itself: how far does it deviate from reality? If this analysis of what makes a case of self-deception paradigmatic is right, then reflection is at best superfluous to understanding how paradigmatic cases function. Why insist that reflection is at the heart of self-deception if we can fully characterize the phenomenon without it and so avoid the associated costs of reflectivism? Remember that on a broadly Strawsonian theory of responsibility we can hold a self-deceived agent responsible without making any claims about how and whether reflection is involved. Whatever the reasons that Scott-Kakures and others believe agency must be reflective, they should be reminded that giving up reflectivism and reflective agency does not entail giving up on moral responsibility.

One final matter: Scott-Kakures motivates his theory with the intuition that Bonnie the cat could not possibly be self-deceived. He believes that people like Mele doom their own theories by being so permissive and that an account of self-deception based on reflective agency solves the problem. While the idea that animals could be self-deceived is indeed strange at first blush, it's important to note that Mele wouldn't be bothered by this criticism.

[10] Though Scott-Kakures would surely want to describe the kind of effort required as reflective, we need not take this tack. As I have sketched it, paradigmatic self-deception requires time and energy to remind oneself of favorable evidence and avoid that which is unfavorable, and neither of these is a (necessarily) psychological notion.

His theory is *deflationary*, after all, and so he's trying to characterize self-deception as a broad phenomenon with minimal requirements. Scott-Kakures and I, by contrast, are much more concerned with explaining what makes for a *paradigm* case of self-deception. This is a different project, and objecting to Mele on the grounds that his theory doesn't "really" capture the phenomenon threatens to devolve into tedious semantic bickering about definitions. Better to avoid that debate entirely, I think: it is impossible for Bonnie to feature in a paradigm case of self-deception, but I certainly won't begrudge a deflationist theory that says she is in *some* sense self-deceived.

2.9 What's The Solution?

I have claimed that Scott-Kakures' claims about the importance of reflection to self-deception is mistaken – or at least that he makes unnecessary and controversial theoretical assumptions. Where Mele and deflationists are too permissive about characterizing self-deception, Scott-Kakures is too demanding. So now the burden is on me to provide an alternative explanation. How can we claim that moral responsibility applies to paradigmatic self-deception if we want to avoid talk of reflective agency? In Chapter III, I will discuss how paradigmatic self-deception is plausibly *intentional* and lay the groundwork for the account of responsibility to come in Chapter IV.

Chapter III
Intentional Ignorance

The small town of Le Roy, New York, was the site of an extraordinary series of events beginning in August 2011. Over the course of that school year, 20 local residents, mostly high school students, suffered from severe physical symptoms including uncontrollable twitching and assorted verbal tics. Susan Dominus (2012) describes the initial outbreak in her article for *The New York Times Magazine*:

> Katie Krautwurst, a high-school cheerleader from Le Roy, N.Y., woke up from a nap. Instantly, she knew something was wrong. Her chin was jutting forward uncontrollably and her face was contracting into spasms. She was still twitching a few weeks later when her best friend, Thera Sanchez, captain of one of the school's cheerleading squads, awoke from a nap stuttering and then later started twitching, her arms flailing and head jerking. Two weeks after that, Lydia Parker, also a senior, erupted in tics and arm swings and hums. Then word got around that Chelsey Dumars, another cheerleader, who recently moved to town, was making the same strange noises, the same strange movements, leaving school early on the days she could make it to class at all. The numbers grew – 12, then 16, then 18, in a school of 600 – and as they swelled, the ranks of the sufferers came to include a wider swath of the Le Roy high-school hierarchy: girls who weren't cheerleaders, girls who kept to themselves and had studs in their lips. There was even one boy and an older woman, age 36 (p. 1).

It was clear to their parents that the children were suffering from some kind of illness – but what? Explanations varied from cramping to Tourette's Syndrome, and parents were especially concerned that the water or land might be contaminated. 41 years earlier, a train carrying toxic chemicals derailed in the town and spilled its cargo. Might that be the answer? When a team dispatched by the activist Erin Brockovich to test the soil around the school was denied access by the school's administration, parents revolted; surely this meant that the school had something to hide.

3.1 Explaining Le Roy

The ultimate explanation was no less dramatic. A neurologist who saw many of the affected students eventually reached a diagnosis: a conversion disorder coupled with mass psychogenic illness, or hysteria. A conversion disorder occurs when the body manifests emotional distress as physical symptoms. Dominus notes in her article that all five of the girls she interviewed lacked "stable relationships" with their fathers, and the town's economy was in decline; there was plenty of stress to go around. Psychogenic illness was assigned because so many other people showed similar symptoms after the initial cases; Katie and Thera triggered the illnesses of the other victims, it appears.

Despite the plausibility of the neurologist's diagnosis, some parents continued to seek out alternative explanations. Another neurologist from New Jersey, Rosario Trifiletti, developed his own theory of what was ailing the sick students, telling them that they were suffering from a set of extremely rare complications from streptococcus, a syndrome called PANDAS. Rusty barrels full of unidentified material were discovered near the site of the old train derailment in an area marked "Hazardous Waste," and the EPA later sent in a team to test them. One father remarked, "'With the government, our own health department, collaborating with the school…it's almost just short of a conspiracy…' 'People are getting that thing, like they're trying to hide something'" (Dominus 2012, p. 1). Brockovich continued to insist that there was still testing to be done. Many parents were confused about how their seemingly happy children – they were stressed about certain things, but who isn't? – could be suffering from a psychological malady. Conversion disorder was a strange and uncomfortable diagnosis, in other words, so much that the parents formed false beliefs about the credibility of the diagnosis. "'It's a very hard pill for me to swallow – what are we, living in the 1600s?'" said another frustrated parent (Dominus 2012, p. 1).

But none of these alternatives was ever really plausible. No environmental contamination was demonstrated. It was "almost impossible" that so many people could have been suffering from PANDAS simultaneously, according to the neurologist who first identified the disease (Dominus 2012, p. 1). Moreover, Trifiletti's particular definition of PANDAS was "so vaguely formulated that it was impossible to rule out" – and if it wasn't falsifiable it certainly wasn't a scientific claim (Dominus 2012 p. 1). By the end of the school year, many of the students who sought treatment for conversion disorder were doing much better. *Reuters* reported the testimony of a neurologist whose 15 patients were "80 to 90 percent cured:" "'The vindication for us is that the patients are better. They've got their lives back'" (Gulley 2012, p. 1). Considering the effort involved in sustaining their false beliefs about the reasonableness of alternative diagnoses, the degree to which those beliefs diverged from the best scientific consensus, and how persistent the beliefs were even in the face of evidence, from the discussion in the last chapter we can plausibly conclude the Le Roy parents are good candidates for paradigmatic self-deception.

If we accept that some of the parents of sick children in Le Roy were self-deceived about the causes of the illness, we might then wonder about the appropriateness of holding them morally responsible for this failing. I earlier endorsed a broadly Strawsonian theory of moral responsibility, according to which responsibility is embedded in a general practice of holding each other responsible via reactive attitudes. Those attitudes, like praise and blame, are appropriate or *apt* in cases of paradigmatic self-deception only if certain conditions are met, and so getting clear on what those conditions are is crucial to establishing cases of paradigmatic self-deception as apt for judgments of moral responsibility.

As we saw in Chapter II, the debate about aptness can be framed as a debate about whether paradigmatic self-deception is *agential* – whether it involves exercises of moral agency. I have expressed dissatisfaction with both Alfred Mele's account of self-deception, which seems to leave little room for agency, and also with Dion Scott-Kakures' account, which characterizes self-deception as an exercise of *reflective* agency. Given how fraught with complexity the debate about theories of agency can be, though, in this chapter I want to lay talk of agency aside and instead focus merely on *intention*. In the next chapter, I will offer an account of what justifies holding people responsible – what makes responsibility judgments apt – for cases of paradigmatic self-deception by arguing that paradigmatic self-deception involves a culpable pattern of intentional activity. In particular, I will claim that paradigmatic self-deception involves *culpable ignorance*: following a robust philosophical literature, we can look at paradigmatic self-deception as a phenomenon in which agents are ignorant of the fact that their false beliefs are unjustified and this ignorance is morally culpable because it is the product of a pattern of intentional action; moral responsibility judgments are thus apt. But first we must establish a sense in which paradigmatic self-deception is intentional.

I will argue that paradigmatic self-deception involves intentional action by way of exercise of the capacity to direct attention in ways that are supportive of the false belief and away from evidence that contradicts it.[1] Schematically, my argument will involve three steps: (1) paradigmatic self-deception involves ignorance, since the agent is ignorant of P, where P is true, while believing not-P, which is false; (2) the ignorance is sustained by a pattern of acts and omissions that preserve the false belief at the core of self-deception; and (3) the acts and omissions in paradigmatic self-deception are the product of certain intentions, and in combi-

[1] I gestured in this direction in the last chapter's discussion of effort as a characteristic of paradigmatic self-deception.

nation they make paradigmatic self-deception an intentional activity. Together, these three claims advance the project of establishing paradigmatic self-deception as a distinctive, morally interesting class of cases, continuing the work begun in Chapter II.

3.2 What About Agency?

Recall that Alfred Mele (1997) holds that intentional cases are "remote" from the paradigm of self-deception. I think this view is mistaken, and so does Dion Scott-Kakures, who argues against it by claiming that self-deception involves reflective agency. As I mentioned above, though, I want to avoid the debate about agency here because theory choice is notoriously dependent on stubborn, deep-seated intuitions; I am unlikely to dissuade anyone from favoring a Kantian theory of agency in the space available. My approach will be to talk about intention only, because intention is plausibly a central component of agency and also because we do not need the concept of moral agency to discuss the aptness of responsibility judgments – only the concept of intention. There are many options for a theory of agency that emphasizes intentions, and some are better than others.[2] Rather than endorsing one theory of agency and incurring the wrath of its detractors, instead my aim will be merely to show that paradigmatic self-deception involves patterns of intentional action, patterns that result in culpable ignorance. This is enough to establish paradigmatic cases of self-deception as distinctively interesting with respect to moral responsibility.

[2] Notably the work of Michael Bratman (2000) and, arguably, Kantians like Christine Korsgaard (1996, 2008). Following philosophers like Gary Watson (1975), I would favor a theory that has something to do with values, the commitments that are most central to our lives. Values will be implemented in behavior, necessarily so because they dispose us to act in various ways. Suppose I take Peter Singer's (1972) prescriptions seriously and value helping others in faraway places even to the exclusion of my own comfort. Then we should expect to see me regularly give up fancy dinners and new electronic gadgets in favor of sending money to groups like Oxfam. The claim then is that intentional actions that implement values are instances of agency. Since values influence dispositions to act, their influence will necessarily be exhibited in patterns of action. The same goes for agency, which is relativized to patterns of action rather an individual instances.

3.3 Self-Deception and Ignorance

3.3.1 The Dispositional Model of Belief

Before I can discuss how paradigmatic self-deception involves intentional ignorance, I must be sure that ignorance is even relevant to understanding self-deception. One major worry is that self-deception need not involve any ignorance at all: if it does not, then any discussion of intentions and culpability for ignorance is a waste of time. So we need to establish that ignorance is both conceptually and empirically plausible as a feature of self-deception. Consider the tradition that self-deception involves not just a single false belief P but also its opposite, not-P. This is the *two-beliefs model* of self-deception, and if accurate then in an important sense self-deception does not involve ignorance: the self-deceived agent already *has* the true belief we might otherwise assume she is ignorant of.

The model holds that self-deceived individuals hold beliefs whose contents are *directly* contradictory, such that the agent simultaneously believes both P and not-P. We can distinguish the notion of directly contradictory beliefs from the notion of beliefs that are contradictory only at some inferential distance: if I believe that P and would also believe not-P if I followed the chain of entailments from my current belief, then I do not hold directly contradictory beliefs and am not self-deceived on a two-beliefs account.[3] Given this understanding, we might distinguish several kinds of two-beliefs accounts. For example, we might say that self-deception occurs only when the two beliefs are in *conscious awareness*, or we might allow that

[3] Note that self-deception still involves some inferential remove from one's beliefs: if I am self-deceived in believing both P and not-P, it would still require some further inference to, for instance, form the higher-order belief that I have those two beliefs or that I believe two things that are contradictory. Self-deception involves beliefs that are directly contradictory in the sense that their *contents* contradict one another.

one or both of the beliefs may be *unconscious*. We can further classify accounts based on what the individual is aware of besides the belief: is she aware of whether the beliefs are justified, that the two beliefs are contradictory, or of the evidence for and against each belief? In cases of paradigmatic self-deception, I think there will always be ignorance of one of the beliefs; this is the kind of ignorance that results from omitting to seek out and consider evidence against the false belief. Other kinds of ignorance might also be present to varying degrees, but they are only indirectly related to the omission and so peripheral to the central issue at hand: asking whether self-deception is intentional reduces to asking whether the agent is intentionally ignorant of a belief – the true belief – where this ignorance is the product of an intentional omission.

An appealing way to avoid these problems is to adopt a certain conceptual theory about beliefs: following a standard account from John Heil (2003), I will stipulate that we view beliefs as primarily *dispositional*.[4] That is, beliefs reveal themselves in behavior, such as in a disposition for an individual to avow things that he believes. For instance, if I believe that there is a beer in my fridge then that might prompt me to offer it to a visiting friend of mine who looks particularly thirsty. I view beliefs as dispositions that are causally mediated by real mental states that are often inaccessible to introspection: the dispositions exist because of real mental states that are causally efficacious in certain circumstances, and introspection doesn't typically give us insight into which dispositions we have.

A dispositional model of belief allows room for ignorance in self-deception and also makes some sense of the otherwise puzzling notion of holding directly contradictory beliefs. Believing that P might involve a tendency to deny that not-P even if the individual also believes that not-P holds, and vice versa depending on which belief is in conscious awareness at

[4] See also Schwitzgebel (2002) for a similar account of beliefs.

the moment. Believing both P and not-P need not entail awareness of the contradiction, since believing that one is in a contradictory mental state requires a further inference. It may not be easy to recognize when one has dispositions – that is, *beliefs* – that generate behaviors at odds with one another, even for a person who reflects often on his actions. The dispositions associated with contradictory beliefs may manifest in different contexts: the man who both believes and does not believe that his wife is having an affair may only seem to believe the former when with his friends, who commiserate with him, and the latter when alone and feeling vulnerable.[5] Finally, bringing P into conscious awareness might involve a tendency to "submerge" or otherwise remove from awareness evidence for not-P or against P, particularly if the agent is motivated to believe P over not-P for some reason. All of this is to say that we can readily make the two-beliefs model conceptually coherent and so ignorance is conceptually plausible as a feature of self-deception.

3.3.2 Ignorance and Partitioning

You might worry that the dispositional account makes self-deception trivially easy to explain. If we can hold contradictory beliefs because beliefs are just dispositions that manifest themselves under different conditions, then it is not puzzling how such beliefs are maintained simultaneously. The self-deceived Le Roy parents might sometimes be highly motivated to seek out alternative diagnoses because of the influence of their false belief that the official diagnosis lacks credibility, whereas the opposite belief might lead them to feel guilty about not finding their children a therapist for conversion disorder. Many cases of self-deception, however, seem to involve situations in which *both* beliefs manifest. Imagine that the parents in-

[5] John Doris (2002) cites Mark Johnston (1992) in noting cases in which one disposition can "mask" another if both dispositions have identical eliciting conditions. I could be disposed to both shyness and friendliness and yet never exhibit my friendliness because it is always "trumped" by my shyness. The best way to think about dispositions overall is that what I will be disposed to do in a particular circumstance will depend on my total complement of dispositions (John Heil, personal communication).

stead feel guilty about not entering their children into therapy *precisely when* they are engaged in finding alternative explanations for the diagnosis. In such cases the dispositional account would not be a satisfying explanation of the behavior.

A further burden, then, lies in how contradictory beliefs could be divided, or "partitioned," from one another so that the person can be aware of one but not the other even when both are simultaneously influencing behavior. Alfred Mele offers a general characterization of how psychologists and philosophers have historically tackled this issue: "Partitioning strategies range from relatively modest attempts to distinguish kinds of awareness to the postulation of full-blown sub-agents with their own motives and plans. In general, partitioners accept a strict interpersonal model of self-deception and seek to locate the requisite divisions within a single mind" (1987, p. 3).

Partitioning views trace back to Raphael Demos (1960), who draws a distinction between "simple awareness" of a belief and "awareness together with attending, or noticing" (p. 3). Demos thought that agents could hold contradictory beliefs because they in some cases lack this second kind of awareness.[6] The paper most often cited as evidence for Demos' view comes from Sackeim and Gur (1979), who describe studies in which subjects who wrongly claim that theirs is not the voice coming out of a tape recorder nonetheless display galvanic skin responses (GSR) indicating voice recognition. Sackeim and Gur assert that this is sufficient to confirm the two-beliefs theory because they suppose that subjects believe both that the voice is theirs and that it is not, but Mele denies that GSR tests are necessarily evidence of

[6] David Pugmire (1969) and D. W. Hamlyn (1971) offer refinements of this intuition. Mele also discusses several more "extreme" partitioning views such as that of John King-Farlow (1963), who claims that "a person is quite often usefully looked at...as a *large, loose sort of committee*. There is a most irregularly rotating chairmanship. The members question, warn, praise, and *deceive* each other" (p. 135). Among philosophers, partitioning views are found in the work of Donald Davidson (1982), David Pears (1982), and John Heil (1989, in a criticism of Davidson), although they extend partitioning as a solution to broader problems of irrationality such as *akrasia*.

genuine belief, noting that they could easily measure a “sub-doxastic sensitivity” to voice-recognition.

Quattrone and Tversky (1984) conducted experiments that purport to offer support for Sackeim and Gur’s theory.[7] Quattrone and Tversky gave subjects two trials in which they were asked “to submerge their forearm into a chest of circulating cold water until they could no longer tolerate it.” Before the second trial, subjects pedaled an exercise bike for 1 minute and were told that there were two kinds of hearts: those with “Type 1” hearts “are frequently ill, are prone to heart disease, and have a shorter than average life expectancy” while those with “Type 2” hearts “enjoy good heart health, have a low risk of heart disease, and show a longer than average life expectancy” (p. 240). Following that, half the subjects were told that a Type 1 heart increases tolerance to cold water after exercise while a Type 2 heart decreases it; the other half were told the opposite.

During the trial, most subjects attempted to shift their pain tolerance in the direction that they had been told corresponded with health and longevity, and most denied that they had shifted intentionally. Those who admitted an attempt to shift tolerance overwhelmingly inferred that they were Type 1 (i.e., unhealthy) in post-trial questionnaire, while those who denied a shift mostly inferred they were Type 2 (i.e., healthy). Because the two groups – those who denied shifting and those who admitted it – had no other distinguishing characteristics, the authors conclude that both groups had an awareness of the motivations for their actions: all the subjects who tried to shift their tolerance were motivated by a desire to believe that they had healthy hearts, even though many of them would not admit it. Thus, the authors ar-

[7] More specifically, Quattrone and Tversky are trying to demonstrate a particular kind of self-deception that they term *deceptive diagnosis*.

gue that those who believed that they had not tried to shift their tolerance also secretly believed that "they acted with a target inference in mind" (p. 243).

Though a full empirical defense of partitioning is beyond the scope of this chapter, I believe modern dual-process theories of cognition offer a promising basis for alleviating worries about empirical plausibility. In addition, they provide an alternative defense of the idea of partitioning in general, in case of lingering worries about the attractiveness of the dispositional model of belief. Such theories generally propose that cognition can be divided into two distinct streams. Stanovich and West (2000) use the terms System 1 and System 2. System 1 is "automatic, largely unconscious, and relatively undemanding of computational capacity" (p. 658). System 2, by contrast, typifies "controlled processing" and "analytic intelligence" (p. 658). The extent to which the two streams interact is the subject of much debate, but given how very differently they operate it is intuitive to imagine that System 1 can have mental states (e.g., beliefs and desires) that are separate from and unavailable to System 2 and vice versa.[8]

This leads naturally to a plausible partitioning of beliefs. Suppose that I consciously (System 2) believe that spiders are not scary but unconsciously (System 1) believe that they are. If I wake up in my tent while camping to discover a tarantula on my arm, my automatic

[8] Some dual-process theorists argue that System 1, in virtue of its speed and automaticity, must be independent of belief since computations involving beliefs are supposed to be slow and intensive. The success of this argument depends heavily on what we take beliefs to be. Those who favor a dispositional account of beliefs, as I do, should find it easier to plausibly claim that beliefs can influence automatic processing since there is no reason that the influence of dispositions would be limited to System 2. Moreover, it is already established that the automatic processes relevant to self-deception, those underlying motivated reasoning, are driven by desires; it is unclear why such processes should not also be sensitive to one's beliefs. In the event that further empirical research rules out entirely a connection between beliefs and System 1 cognition, however, I might appeal instead to the distinction between personal and sub-personal cognitive systems originated by Daniel Dennett (1969). As Jennifer Hornsby (2000) explains, "For Dennett, *belief* is predicated in exactly the same sense of a person as of any other, low-grade intentional system" (p. 8). So it makes sense to ascribe beliefs to individual systems, and those beliefs will necessarily be isolated from one another.

reaction may be to panic, with my belief that the tarantula is scary effective via System 1. But after a few moments, my cooler, more analytic reasoning might kick in via System 2 as my belief that the tarantula is not scary becomes effective instead. Because the two cognitive streams that controlled my two very different reactions are not connected, we can intelligibly say that they could hold beliefs that are directly contradictory. Because System 1 is unconscious, I can easily be unaware that the contradiction exists.

More recently, Stanovich (2011) has argued that System 2 can be further subdivided into a "reflective level" and an "algorithmic level" – both are "non-autonomous" in contrast with System 1 – making for a tripartite model of cognition. The reflective level captures individual differences at the level of "higher-level goal states and thinking dispositions," like conscientiousness and diligence, while the algorithmic level captures differences at the level of intelligence, or "the efficiency of functional cognitive machinery that carries out mental tasks" like matching pairs of words or memorizing information (p. 26). If this is right, then self-deception may involve beliefs associated with any of three distinct types of cognition, again supporting the notion that partitioning is a feature native to the mind itself. So there is at least some empirical evidence for partitioning.

3.4 Acts and Omissions

3.4.1 The Act of Self-Deception

Now that I have established that paradigmatic self-deception plausibly involves ignorance, we turn to the pressing matter of showing how that ignorance is intentional. In the last chapter I claimed that paradigmatic self-deception involves effort both to pursue evidence that supports the false belief and avoid evidence that contradicts it. I can now refine that characterization. More specifically, paradigmatic self-deception typically involves both a pattern of *acts* and a pattern of *omissions*, both of which are intentional: self-deceived agents intentionally act to support their false beliefs and intentionally omit to find ways to challenge them, and together the act and the omission produce ignorance of the truth of P while sustaining the false belief that not-P. Let's start with the act: the self-deceived individual continually tries to reinforce the false belief by thinking about evidence that supports it and also seeking out new evidence that is likewise favorable. That is, she acts on an intention to support her belief. I mean for the intentionality to be uncontroversial: the Le Roy parents sought out alternative explanations for the outbreak because they held (false) beliefs about the trustworthiness of the official explanation and because they intentionally sought ways to support those beliefs.

You might think that in cases like Le Roy, this kind of "drowning out" makes it easy to establish that their ignorance was, on balance, intentional, and thus that their self-deception was intentional.[9] The self-deceived parents who resisted the neurologist's diagnosis of conversion disorder can be contrasted with the more reasonable parents who, despite their suspicions, sought psychiatric treatment and whose children greatly improved. In the words of their treating neurologist: "'[N]obody ever got harmed by taking a good hard look at your life'" (Dominus 1). Why couldn't the self-deceived parents realize this? Instead they formed false beliefs about the diagnosing neurologist's credibility and devoted their resources not to helping their children get better but to trying to find ways to show that the diagnosis was mistaken. We might even infer that seeking alternative explanations was a tactic to increase their

[9] The parents of the sick Le Roy children make an interesting example here, especially as compared to the Romney campaign example from the last chapter. Romney's self-deception was sustained mostly by avoiding evidence that might have led him to revise his false belief. He and his staff believed at the start of their polling that turnout would be demographically favorable, and they didn't spend time gathering additional evidence for that belief. The Le Roy parents, by contrast, sustained their false belief primarily by seeking out additional evidence to support it. That is, they didn't avoid evidence against their false belief so much as drown it out by assembling evidence that made the false belief seem plausible.

confidence that the diagnosis lacked credibility and so make the false belief easier to sustain. These parents' energy and resources were finite – they all had busy lives that were probably more stressful than average – so why expend so much effort if not to make it easier to perpetuate their self-deception?

If this sketch of events is accurate, we might say that the Le Roy parents intentionally deceived themselves by intentionally acting to find alternatives to the conversion disorder diagnosis in a way that produced ignorance of the fact that the conversion disorder diagnosis was quite reasonable.[10] This explanation reduces Le Roy parents' self-deception to a pattern of intentional action, and so we might think we have described how their self-deception was intentional.

3.4.2 The Omission of Self-Deception

Yet this is too quick. The Le Roy parents might have intentionally sought out alternative explanations for their children's illness, but at the same time they also seem to have *omitted* to consider seriously the evidence that favored the conversion disorder diagnosis.[11] Even if we think that the act, rather than the omission, was doing most of the work in sustaining the false belief, the omission played a necessary role in their being ignorant of the truth of P while sustaining the false belief that not-P. The ignorance could not have been sustained

[10] The qualifier at the beginning of this sentence is meant seriously. There are real concerns about distorting what actually happened, though many of them are due simply to not having complete information. My aim in describing the Le Roy case as well as all the others discussed elsewhere is to present a plausible gloss on the agents' mental states and argue from there to claims about intention and paradigmatic self-deception. Some of the details will end up being stipulative, but this is inevitable and real-world cases are, on the whole, much more interesting to talk about than pure thought experiments (which are stipulative in every way).

[11] For the sake of simplicity, I characterize omissions as merely "failing to consider" the relevant evidence. One wrinkle is that self-deceived individuals may consider the evidence but not in the "right" way – the way that would lead them to jettison their false beliefs.

without the omission, so showing that paradigmatic self-deception is intentional requires showing that the pattern of omissions – and not just the pattern of acts – is intentional.

Even if it is fairly obvious that the pattern of acts was intentional, as could be the case with the Le Roy parents, we cannot establish that their self-deception was intentional without succeeding at the difficult task of showing that the pattern of omissions, too, was intentional. Moreover, it needs to be clear how the pattern of intentional acts and the pattern of intentional omissions are *related*. I claim that they are both part of the activity of being paradigmatically self-deceived, so the intentions in the acts and omissions must be connected somehow. Getting clear on that connection is crucial to showing that the *combination* of acts and omissions – paradigmatic self-deception – is intentional. This is the project of the rest of the chapter.

3.5 Intentional Content

The problem of how acts and omissions are related in paradigmatic self-deception comes down to questions about what the *content* of the intention is in both cases. Naively we might think that intentionally doing X requires having an intention with the content "DO X." On such an account, paradigmatic self-deception could seemingly never be intentional, since intentionally deceiving myself would require that I have the intention "DECEIVE MYSELF." This would be bizarre. As Mele (1997) has noted, such an intention seems to be self-undermining: "It is hard to imagine how one person can deceive another into believing that *p* if the latter person knows exactly what the former is up to. And it is difficult to see how the trick can be any easier when the intending deceiver and the intended victim are the same person" (p. 92).

Mental partitioning strategies might offer a way around the problem, since we could suppose that the intention really does have the content "DECEIVE MYSELF" and that it is hidden from conscious awareness in some partition that is inaccessible to introspection. While this would make paradigmatic self-deception intentional, it could hardly provide any grounds for moral responsibility: in what sense could I be morally responsible for acting on an intention over which I have no conscious access or control? Justifying the aptness of responsibility judgments seems ruled out by partitioning. We need to find an intention that suffices to make the acts and omissions intentional while allowing for the possibility of culpability; I thus want to avoid any view that requires that the intention be hidden from conscious awareness.

The solution to this problem I favor involves rejecting the simple account of intentional action in favor of an account that says that I can intentionally do X without having an intention with content "DO X." How does this work in paradigmatic self-deception, then? One proposal is that acts and omissions in paradigmatic self-deception are intentional because they are *side effects* of some other intentional action. Consider a famous case from Michael Bratman (1984) in which a marathon runner runs a race and in so doing wears down his shoes. The runner did not intend – that is, possess the intention – to wear down his shoes, but Bratman rightly notes that he could be said to have intentionally worn them down nonetheless. Similarly, Gilbert Harman (1976) offers the case of a sniper who intentionally alerts the enemy to his presence when he fires at a soldier even though he does not *intend* to alert the enemy.[12] We could then claim that in paradigmatic self-deception the individual intentionally acts to find evidence that favors the false belief and omits to find evidence against it because the pattern of omissions is a side effect of some other (pattern of) intentional action.

This approach immediately runs into problems, however. First, it's not obvious what the relevant intentional action would be. The most natural thing to say is that we intend to deceive ourselves and as a side effect intentionally act and omit in the relevant ways, but this

[12] I credit Mele (1992) as well as Mele and Sverdlik (1996) for bringing these examples to my attention.

begs the question about intentionality that we set out answer at the beginning. We must also take seriously the details of Bratman's and Harman's examples. Bratman is confident that the marathon runner intentionally runs down his shoes only when he builds in the assumption that the runner *believes* that he will wear them down, and Harman assumes that the sniper believes that the gain from killing the soldier is worth the risk; this is just to say that the side effects are *foreseen*. It is implausible that the paradigmatically self-deceived individual occurrently believes – foresees – that performing some action will result in a pattern of acting and omitting; that kind of foresight would be self-undermining just as on the simple proposal. Unless such a belief exists, it is hard to believe that paradigmatic self-deception is intentional because it is a side effect of some intentional action.[13]

Randolph Clarke (2010) offers a better solution: whether an action is intentional seems to depend on its having certain *relevant* content, where the notion of relevance is vague: "There are cases in which the agent clearly has a relevant intention, cases in which the agent clearly does not, and cases in which it is unclear whether the content of any intention possessed by the agent is suitably relevant" (p. 166). A clear example is trying intentions: Charles intentionally and successfully quits smoking by forming the intention to *try* not to smoke. This is not enough to explain why acts and omissions in paradigmatic self-deception are intentional, though; the Romney campaign staff did not intentionally omit to find evidence against their false belief merely by forming the intention to try to omit to find that evidence. But trying intentions are a specific example of a more general principle, that the content of an intention is often a *means* to do whatever we are intentionally doing. Charles' intention contains the implementation plan for his intentional omitting to smoke, and so too the Romney

[13] It might be possible to argue that side effects can be intentional if they are merely *foreseeable*, but this would be a significant undertaking.

campaign's intention perhaps had content that served as a means of intentionally omitting to find evidence against the campaign's false belief.

What might be the intentions at work in the case of intentional acts and omissions in paradigmatic self-deception? I tentatively propose that self-deceived agents in such cases intentionally act to support their false belief with the intention "DIRECT MY ATTENTION TOWARDS EVIDENCE THAT SUPPORTS MY BELIEF" and intentionally omit to challenge their false belief by acting on the intention "DIRECT MY ATTENTION AWAY FROM SITUATIONS, PEOPLE, AND OTHER SOURCES OF INFORMATION THAT MIGHT THREATEN MY BELIEF." Yet this is still not enough. We need to know more about how these two intentions are relevant (in Clarke's technical sense) to the overall activity of being self-deceived. I agree with Clarke that relevance is vague and so necessary and sufficient conditions for relevance do not exist, but merely stipulating that the two intentions suffice to make self-deception intentional is surely unsatisfying.

Before moving on, I should also concede that an account of intentional action based on the general notion of relevant content, rather than on specific, narrowly defined types of intentions, is unconventional. Mele (2001) contends that intentional cases of self-deception are remote from the paradigm primarily because he thinks that self-deception could be intentional only if the self-deceived person has an intention "to deceive herself into believing [the false] hypothesis, or to cause herself to believe this, or to make it easier for herself to believe this" (p. 18). I claim that self-deception can be intentional even if the individual has none of these particular intentions. While this may at first blush seem odd, the intentions Mele names are just a few among potentially many intentions that have content relevant to self-deception. Once you accept that a person can intentionally deceive himself without having an intention with content "DECEIVE MYSELF," it remains only to argue about what kinds of intentional content are permitted. Giving necessary and sufficient conditions for what counts as relevant content is a fool's errand, but I think that by restricting the discussion to the particular case of paradigmatic self-deception we can avoid that problem and understand how a person could intentionally deceive herself without having any of the intentions Mele names. I will make no claims about what counts as relevant content in general, only what plausibly counts as relevant content in self-deception.

3.6 Attentional Control

My task now is to explain how self-deception is intentional as a pattern of controlled exercises of attention that operate under an intention relevant to self-deception. As background, it is important to understand how psychologists treat attentional control: to what extent, and by what means, do humans exert control over attention? That we do to some degree is undeniable, as philosophers of perception like Ned Block (2010) have discussed. Consider the famous Stroop effect: reading the names of colors when they are printed in a color other than the one they name takes more effort than if they are printed in the same color. As Posner and Snyder (1975) observe, "[O]ne intends to avoid reading the words, but it is not possible to do so completely" (p. 206). The Stroop effect shows that our control over attention is not total, but it also shows that we can – with difficulty – ignore the distracting stimulus and successfully read the word. But how are controlled exercises of attention like this accomplished?

Some important clues come from studying how attention develops during childhood. Rueda, Posner, and Rothbart (2004) explain that one of the first signs of attentional control is the ability regulate emotional distress: "In infants as young as 3 months, we have found that orienting to a visual stimulus provided by the experimenter produces powerful, if only temporary, soothing of distress. One of the major accomplishments of the first few years is for infants to develop the means to achieve this regulation on their own" (p. 288). More sophisticated control is revealed in the ability to solve attention-related conflicts:

> [I]nfants develop the ability to resolve conflict between line of sight and line of reach when retrieving an object. At 9 months, line of sight dominates completely. If the open side of a box is not in line with the side in view, the infant will withdraw its hand and reach directly along the line of sight, striking the closed side. In contrast, 12-month old infants can simultaneously look at a closed side and reach through the open end to retrieve a toy (RPR 2004, p. 288)

By around 30 months, toddlers can perform more complicated spatial tasks such as the following: "Children sat in front of two response keys, one located to their left and one to their right. Each key displayed a picture, and on every trial a picture identical to one of the pair appeared on either the left or the right side of the screen. Children were rewarded for responding to the identity of the stimulus, regardless of its spatial compatibility with the matching response key" (RPR 2004, p. 289). Because the test produces conflicts between identity and spatial position, the toddler subjects are required to expend significant effort to pick the right answers. This effort manifests on fMRI as activation of the anterior cingulate gyrus, part of what the authors term the "executive attention network" (RPR 2004, p. 287).

But perhaps these kinds of attentional control are too weak to be heavily implicated in self-deception. In early childhood we come to be able effortfully resolve simple attentional problems to allow us to navigate our environment in ways we could not, but that is not very close to what goes on in self-deception. Surely self-deception requires control that is robust, if the self-deceiver is to sustain a false belief over a long period of time and despite copious countervailing evidence. Consider the phenomenon of *inattentional blindness*, wherein subjects attend to part of a scene so much that they miss out on another part that would otherwise be obvious. Inattentional blindness is especially interesting in thinking about self-deception because it shows how controlled attention can sometimes lead us astray. Simons and Chabris (1999) give an illustration:

> Perhaps you have had the following experience: you are searching for an open seat in a crowded movie theater. After scanning for several minutes, you eventually spot one and sit down. The next day, your friends ask why you ignored them at the theater. They were waving at you, and you looked right at them but did not see them…We feel that we perceive and remember everything around us, and we take the occasional blindness to visual details to be an unusual exception (p. 1059).

The big idea is that our feelings of noticing and recording everything around us are often quite mistaken; much of what represent and remember depends on what specific features we attend to. Neisser (1979) famously demonstrated inattentional blindness with an experiment in which subjects were so wrapped up in watching a basketball game and counting the number of passes made by a team that they failed to notice four seconds of superimposed video of a woman with an umbrella walking across the screen. But perhaps the woman was just too hard to notice, only appearing for a short time and then only faintly. In a now-classic paper, Simons and Chabris (1999) updated the experiment with tests in which a man with an umbrella or a man in a gorilla suit was part of the scene itself for as long as nine seconds. Only 50% of subjects noticed the gorilla man, even when he stopped in the middle of the frame to beat his chest! Those who missed it were shocked when the video was replayed with explanation.

Inattentional blindness can be modulated by specific features of the scenario. For example, Steven Most and colleagues (2001) argue that the likelihood of inattentional blindness can depend on the unexpected object being different from the attended objects and similar to the ignored objects in the scene. They also note, tantalizingly, that the similarity effect "may be driven partially, if not largely, by the selective ignoring of irrelevant stimuli" that are not in the subject's "attentional set" (p. 16).

The notion of an attentional set, which is ubiquitous in the attention literature, comes from Folk, Remington, and Johnston (1992). They argue against the view that shifts in visual spatial attention "are the result of relatively inflexible, 'hard-wired' mechanisms triggered by specific stimulus properties, namely abrupt luminance changes" (p. 1041). Instead, they claim:

> [Attention] can be 'configured' or 'set' to respond selectively…[A]ny particular system configuration, or 'attentional control setting,' is assumed to be a function of current behavioral goals. (In the absence of any current task goals, the system might default to settings based on long-term biases.) With a control setting established, events exhibiting the critical properties will involuntarily summon attention…Stimuli not exhibiting these properties will not involuntarily summon attention (p. 1041).

Attentional sets have been studied by neuroscientists as well. Banish and collaborators (2000) found that establishing and maintaining an attentional set involves the prefrontal cortex (in particular the DLPFC), the locus of executive control, with activation levels depending on how difficult the set is to impose.

Attentional sets tie in to the broad distinction psychologists draw between *endogenous* and *exogenous* attention (Posner & Snyder 1975). Endogenous attention, which William James (1890) called *active* attention, is attention directed by "the intentions or strategies of the observer" (Ruz & Lupianez 2002, p. 283). That is, endogenous attention is the kind we use to achieve a goal. Exogenous attention, by contrast, is activated when "the properties of the environment…attract the attention of the observer independent of his/her intentions" (Ruz & Lupianez 2002, p. 283). For example, a loud, unexpected noise typically commands attention even if the observer is engaged in another activity. The distinction is important because attentional sets help us understand the means by which endogenous attention is implemented.[14] Naively we might think that endogenous attention can involve arbitrarily fine-grained control over which features of the environment we attend to. But this is mistaken, since endogenous attention merely programs the content of the attentional sets, which then operate automatically to summon attention on stimuli that match their programming. At most, say Posner and Snyder (2004), "[S]ubjects can program their conscious attention to (1) receive input from a particular input channel or area of memory and (2) perform particular operations upon received information" (p. 216). Thus attentional sets place important limits on endogenous – conscious – control over attention.

So an attentional set is a collection of settings for the visual attention system under which the system operates to serve the goals of the moment. If something like this view is right, it has important implications for self-deception. First note that cases of paradigmatic self-deception will involve endogenous exercises of attention, since they are goal-oriented. In paradigmatic self-deception the agent has the specific goals of finding evidence that supports the false belief and omitting to find evidence that contradicts it, so those goals will continually structure the agent's attentional processes in ways designed to facilitate success. This model aligns with the work of Peter Gollwitzer (1999), a psychologist who has has extensively studied the ways in which goals structure intentional behavior. Gollwitzer distinguishes between *goal* intentions, "end states an individual wants to attain," and *implementation* intentions, which "create a mental link between a selected cue or situation and a goal-directed response," thus committing the agent "to perform this goal-directed response as soon as the specific sit-

[14] The existence of attentional sets may also partially collapse the distinction between endogenous and exogenous attention. If all attention is truly mediated by attentional sets, then there are no truly exogenous attentional shifts; any putatively exogenous shift occurs via the activation of some attentional set (Folk, Remington, & Johnson 1992).

uation is encountered" (Achtziger, Gollwitzer, & Sheeran 2008, p. 381). We can thus identify the paradigmatically self-deceived person's goal intentions to find and omit to find relevant evidence and her implementation intentions to direct her implementation intentions in ways that facilitate achieving these goal intentions. This should have huge implications for the way the agent processes visual experiences, since a large amount of the evidence relevant to any particular case of paradigmatic self-deception is surely encountered visually. When I say that the self-deceived agent looks at the world differently from a normal person, I mean it quite literally.

We can also say that those endogenous exercises of attention will ultimately be implemented as the programming of certain attentional sets at some very low level. In the case of self-deception it's hard to say what the content of those attentional sets will be. But this is not too concerning: if we know that self-deception operates via endogenous exercises of attention, then we have a guarantee that there will be some collection of attentional sets that implements it. The important point is that implementing the attentional strategies that sustain paradigmatic self-deception reduces to tapping into certain attentional sets. They are thus among the building blocks of self-deception.

3.7 Mindfulness and Relevant Intentions

The foregoing discussion has gone some of the way towards explaining how attentional control can be used to facilitate self-deception. But there are still many questions. The literature up to now has focused mostly on visual attention, which is surely part of sustaining the false belief but is nowhere near the whole story; self-deception isn't merely an exercise in selectively processing occurrent visual stimuli, not least because in paradigmatic cases self-deception is diachronic and not merely synchronic. More pressingly, I still need to show how paradigmatic self-deception's attentional activities of acting and omitting can be structured by an overarching intention with content *relevant* to self-deception.

To answer these questions, let's look at a concrete example of how seemingly disparate intentions can be unified in a complex, diachronic, intentional activity that principally involves attention. Recent research into meditation and what psychologists call *mindfulness* offers both an example of relevant intentions in practice as well as further insight into the specific attentional mechanisms that might underlie self-deception itself. Mindfulness meditation is a therapeutic approach descended from Buddhism that seeks to treat psychological problems such as anxiety and depression by refocusing the practitioner's attention away from stressors. The psychologist Scott Bishop and colleagues (2004) offer a two-part definition of mindfulness:

> The first component involves the self-regulation of attention so that it is maintained on immediate experience, thereby allowing for increased recognition of mental events in the present moment. The second component involves adopting a particular orientation towards one's experiences in the present moment, an orientation that is characterized by curiosity, openness, and acceptance (p. 232).

In their view, practicing mindfulness involves developing three distinct skills. First, practitioners attend to their breathing "so that thoughts, feelings, and sensations can be detected as they arise in the stream of consciousness" (p. 232). This requires a skill that Bishop et al call *sustained attention*: practitioners gain facility with focusing for extended periods on what they experience in the moment, instead of letting their thoughts wander away from present experience. When a thought, feeling, or sensation is encountered while practitioners are focused on their breath, they deploy a second skill, *switching*, to direct attention back to the breath and continue the meditation. Switching is thus a skill of flexibility in attention: practitioners develop facility with consciously redirecting their focus. The third skill is developing a *non-elaborative awareness* of one's thoughts, feelings, and sensations. Rather than elaborating – that is, ruminating – on their mental states, practitioners are able to acknowledge them without engaging with them further, instead returning their attention to the breath.

Coffman, Dimidjian, and Baer (2006) describe a mindfulness therapy for depression. Mindfulness is an especially promising method of treatment because depression sufferers are at high risk for relapse even after seemingly successful treatment, and Coffman and colleagues propose this is because "individuals who have experienced one or more depressive episodes have developed associations between sadness and negative thought patterns, and therefore they differ from never-depressed individuals in the negative patterns of thinking triggered by the ordinary sad moods that are unavoidably part of life" (pp. 32-33). Depressed individuals are also highly likely to engage in rumination, perhaps even believing that "dwelling on their problems and trying to understand their inadequacies will yield important insights about their moods and how to improve them" (p. 33).[15] Mindfulness therapy for depression involves teaching the skills that Bishop and colleagues describe to "allow individuals, in times of sadness, to interrupt their old habitual patterns of thinking or behaving so that these moods remain mild or transient and do not escalate into more serious affective states" (p. 35).

Empirical evidence on this therapy is encouraging. A study by Teasdale and colleagues (2000) treated 145 recently depressed patients for eight weeks with a combination of group training sessions in mindfulness techniques and also awareness exercises assigned each week as homework. The treatment program aimed to help patients "cultivate an open and acceptant mode of response, in which they intentionally face and move in to difficulties and dis-

[15] This surely fails to apply in many cases of depression, but mindfulness therapy may be especially helpful in cases where it does.

comfort, and to develop a decentered perspective on thoughts and feelings, in which these are viewed as passing events in the mind" (p. 618). The experimenters illustrated the goal to the patients with an analogy: we often drive in "automatic pilot," being "unaware of the road or other vehicles, preoccupied with planning future activities or ruminating on a current concern" (p. 618). Patients were instead encouraged to emulate "mindful" driving, which emphasizes "early detection of relapse-related patterns of negative thinking, feelings, and body sensations, thus allowing them to be 'nipped in the bud'" (p. 618). Results were significant. For the worst sufferers, the 77% of subjects with a history of at least three depressive episodes, the risk of relapse was nearly halved compared with standard treatments.

Neuroscientific evidence on mindfulness is also promising. Slagter and colleagues (2007) found that subjects with three months of training in mindfulness techniques displayed reduced elaborative processing of target stimuli in an "attentional blink" test measured by placing electrodes on subjects' scalps. As they explain, "When two targets (T1 and T2) embedded in a rapid stream of events are presented in close temporal proximity, the second target is often not seen. This deficit is believed to result from competition between the two targets for limited attentional resources" (p. 1228). Success on the test requires subjects to allocate attentional resources efficiently to T1 so that they will also have resources available for T2, and mindfulness practitioners performed much better than average because they were able to dampen the "ongoing mental noise" that would otherwise interfere with the ability to attend to T2 (p. 1234). The authors conclude that their results confirm "first-person reports that this style of meditation affects attentional processes and can significantly affect the way stimuli are processed and perceived" (p. 1233). Results like these are especially interesting, as Lutz and collaborators (2008) note, because they indicate that mindfulness is a skill that can be effective even when practitioners are not currently meditating:

> Because participants were not engaged in formal meditation during task performance, these results provide support for the idea that one effect of an intensive training in [mindfulness] meditation might be reduction in the propensity to 'get stuck' on a target, as reflected in less elaborate stimulus processing and the development of efficient mechanisms to engage and then disengage from target stimuli in response to task demands (p. 166).

What conclusions for self-deception might we draw from this literature? First, the efficacy of mindfulness training provides strong evidence for intentional control over attention: practitioners make a conscious effort to keep their attention focused on breathing and not on rumination, an exercise that is quintessentially intentional. Paradigmatic self-deception, as a pattern of acts and omissions, is largely an exercise of directing attention in certain ways, namely towards evidence that favors the false belief and away from evidence that threatens it. With mindfulness we have a compelling example of how even sophisticated exercises of attention like those at issue in paradigmatic self-deception could be intentional.

As I conceive of it, sustaining a false belief in paradigmatic self-deception requires skills very much like those involved in mindfulness except that they are turned to different ends. Sustained attention could be implicated in self-deceived individuals' ability to focus intently on evidence that favors their false beliefs – instead of breathing as in mindfulness, the target of focus in paradigmatic self-deception is the most vivid evidence for the false belief – while switching could explain how individuals are able to quickly refocus on the evidence favoring the false belief when they encounter something that threatens it. Non-elaborative awareness seems directly implicated in the activity of omission: whenever self-deceived individuals encounter evidence that threatens the false belief, they are able to avoid the rumination that might lead them to take that evidence seriously. If this model is right, then individual differences in these skills would predict the ease with which we can be paradigmatically self-deceived.

Second, mindfulness seems to be an example of how someone could intentionally recover from a disorder like depression without necessarily intending to recover. Imagine a person who is deeply depressed and wishes desperately that she could feel better. A friend persistently encourages her to try mindfulness techniques. She is skeptical but goes along with it by way of humoring her friend. The therapy ends up working spectacularly well for her, and in a matter of months she is feeling better than she has in years. She never formed the intention to recover from her depression – she had never even considered that it would work, so it wasn't even an intention to *try* to recover – but instead formed the intention to engage in a series of complex, intentional activities as part of an ultimately effective therapy. That intention is *relevant* to a desire that she has, namely to get better. In this case, I think it makes sense to say that her recovery from depression was intentional despite not her not intending to recover. She possessed and acted on intentions that effectively treated her depression, and those intentions were relevant to a desire she possessed, namely to not be depressed, and thus relevant to her overall intentional activity of treating her depression.

I can now explain the sense in which paradigmatic self-deception is intentional. I claim that in paradigmatic self-deception individuals intentionally deceive themselves in that they (1) intend to engage in activities that sustain their false belief that P, in particular the activities of acting to direct their attention towards evidence that favors P and omitting to direct their attention towards evidence that threatens it; (2) act on those intentions; and (3) possess a desire to believe that P rather than not-P that makes their intentions *relevant* to self-deception, since that very desire is what led them to acquire in the first place, via motivated reasoning, the belief that P.

Thus paradigmatically self-deceived individuals intentionally deceive themselves despite not having the intention to deceive themselves. Rather, they act on intentions that are effective in sustaining their false belief and that are relevant to self-deception in that these individuals possess a desire to continue to believe that P rather than not-P. The proposal, then, ends up being not all that dissimilar to what Bratman and Harman suggested. They and I both think that doing something intentionally while not intending to do it requires having some relevant mental state. For them it is a belief, and for me it is a desire.

3.8 Conclusion

On the view that I have been articulating, self-deception is a kind of *failure*: paradigmatically self-deceived individuals fail to seek or consider evidence that would threaten their false beliefs. This is not surprising even on existing views such as Mele's, since like me he thinks that cases of self-deception testify to the many interesting and exotic ways that reasoning can go wrong. But unique to my account is the claim that self-deception is also an *achievement*. Being self-deceived, especially in the most spectacular cases, turns out to require the skillful, intentional deployment of several different kinds of skills – hardly the kind of thing that one blunders into accidentally.

With that intuition and the arguments of this chapter in mind, we can move forward with the project of establishing a basis for holding individuals morally responsible for being self-deceived. As we will see, intentional self-deception involves a kind of negligence for which paradigmatically self-deceived are typically morally responsible. Exploring negligence and its role in undergirding the moral responsibility of paradigmatic self-deception is the task of the next chapter.

Chapter IV
Negligence and Responsibility

Having established how paradigmatic self-deception can be intentional, it remains to show how it could be culpable – how we can ground our reactive attitudes of holding self-deceived individuals morally responsible for their actions. It seems appropriate for Romney's supporters to, for instance, blame and resent the campaign for ignoring the evidence that threatened its misguided polling model. The challenge is to show that those attitudes are warranted. That Romney intentionally ignored the evidence is not enough to get us to culpability: many actions are intentional without being culpable, and we may be culpable for actions that are unintentional. I aim to show that the specific features of paradigmatic self-deception make it a plausible candidate for culpability.

4.1 Culpability and Responsibility

First, though, I should distinguish between the concept of moral culpability and the concept of moral responsibility; while some philosophers seem to use the terms interchangeably, on my account they are distinct concepts. I stipulate that moral responsibility applies to an act whenever it is apt for the reactive attitudes. The reactive attitudes can have either a positive or negative valence – praising versus resenting, for instance – and which is appropriate depends on the circumstances. Thus far I have been concerned with cases of paradigmatic self-deception in which negative attitudes will be appropriate: I claim that we can reasonably blame or resent the Romney campaign for its actions in sustaining its false beliefs, and the same for the Le Roy parents. Whenever negatively valenced reactive attitudes are apt for an act, I say that act is *culpable*. So culpability obtains whenever the act is apt for reactive attitudes – whenever moral responsibility applies – and when those particular attitudes will have a negative valence. Any argument that successfully establishes that paradigmatic self-deception can be culpable will also establish that it is morally responsible, because culpability presupposes responsibility.

4.1.1 Excusing Conditions

Any case in which the conditions for moral responsibility are satisfied – that is, any case some kind of reactive attitude is apt, whether positive or negative – and yet the act in question is *not* culpable can be explained by the presence of certain *excusing conditions*. If I intentionally drive recklessly and endanger pedestrians, I am typically culpable for having done so. But if I was acting under the influence of a mad scientist's mind control ray, then culpability will not apply despite my acting in ways that generally entail culpability. Many and perhaps even most cases of paradigmatic self-deception will be culpable, I think, but some certainly will not be. Consider the psychological literature on positive illusions in romantic relationships. Partners tend to idealize each other and the relationship in ways that are conducive to its long term success. Sandra Murray (1999) explains:

> [H]appily married couples typically attribute their spouses' negative behaviors to specific, unstable features of the situation rather than make more threatening attributions to dispositional weaknesses. Satisfied, secure dating intimates also appear to protect their commitments by defensively misinterpreting their partners' possible attraction to others. Similarly, satisfied individuals are unlikely to think about or even look at attractive alternative partners, and they derogate available partners in efforts to support idealized views of their own partners (p. 24).

The behaviors described will plausibly involve holding false beliefs, and activities like avoiding thoughts of other potential partners fit exactly the pattern of behavior I have claimed is characteristic of paradigmatic self-deception. But though the conditions for moral responsibility may be satisfied, in cases like these the partners' actions are almost certainly not culpable. Insofar as sustaining their false beliefs explained social norm of partners committing fully to each other and to the success of the relationship, that excuses what would otherwise be a culpable behavior. So romantic relationships are fertile ground for non-culpable self-deception.

Another vivid example comes from Simon Keller, who in *The Limits of Loyalty* (2007) claims that being a good friend requires holding certain false beliefs that are good for the relationship. Suppose Joey asks Chandler whether his next acting audition will lead to his big break and Chandler unreflectively says "No." Then, argues Keller,

> [t]here is a crisis in the friendship because Chandler does not believe that this acting job will be Joey's big break. If he were really a good friend, Chandler would have more optimistic beliefs about Joey's prospects – whether the evidence supported those beliefs or not….Sometimes, to be a good friend, you need to compromise your epistemic integrity (pp. 24-5).

If Keller is right that the norms of friendship require a degree of looseness with the facts, then being an excellent friend might lead to a case of non-culpable paradigmatic self-deception. But once again, culpability is ruled out only because of the presence of excusing conditions. For the rest of the chapter I will discuss how and why paradigmatic self-deception is culpable. I claim that paradigmatic self-deception is typically culpable, absent any excusing conditions like those we've seen here.[16]

4.1.2 Moral and Epistemic Responsibility

We also need to be careful to respect the distinction between epistemic and moral varieties of culpability. The difference between the epistemic and the moral is most obvious if

[16] For any particular case, the burden in showing that paradigmatic self-deception is not culpable will be in identifying the excusing conditions and explaining how they excuse. Absent such an explanation, the self-deception is culpable.

we suppose that there independent systems of norms that govern epistemic and moral behavior. Epistemic norms, then, are those that prescribe certain ways of gathering and responding to evidence – in particular, those ways that are generally truth-conducive. Moral norms, by contrast, prescribe acting in accordance with whatever morality requires.[17] Assume for the moment that a consequentialist theory of morality is correct. Then, if a case of paradigmatic self-deception turns out not to have bad consequences, it will not violate any moral norms and will thus not be apt for judgments of moral responsibility. Instead that case of self-deception will be at most epistemically, not morally, culpable. Suppose that the Le Roy parents' self-deception had had no bad consequences: once the media attention died down, all the sick children in Le Roy began to feel better in short order all on their own, with no need for any therapy. If things had gone that way, the self-deceived parents would be epistemically culpable, since they did not believe what the evidence warranted, but they would seemingly not be morally culpable, since their having a false belief didn't produce any morally bad effects. If this story is correct, then there is a problem with the task I have set for myself in this chapter: in at least some cases of paradigmatic self-deception, judgments of moral responsibility will not be apt, contrary to what I claim.

I have two responses to this worry. The first is to note that much depends on what, exactly, the agent is responsible *for*. Discussions of epistemic responsibility are preoccupied with whether an agent is responsible for what he *believes*: Is Mitt Romney criticizable for wrongly believing that his campaign's turnout model was accurate? Are the Le Roy parents criticizable for wrongly believing that the conversion disorder diagnosis was unreasonable? The answers to these questions depend on whether epistemic norms were violated, which in

[17] On an expressivist metaethics what morality requires *just is* whatever the norms of a given society are, but let's put that wrinkle aside.

both cases they quite plausibly were. But of course this is not enough to establish *moral* culpability, since we are supposing that moral and epistemic norms are distinct. How do we get to moral culpability, when we seem to have only epistemic culpability? Based on the account of paradigmatic self-deception I have been developing, a solution presents itself: I have argued that moral responsibility for self-deception derives not from the false beliefs that the agent has but from the *actions* she takes to sustain them. And moral responsibility attaches to actions, with the degree of culpability depending on the degree to which the moral norms in question were violated. Thus paradigmatic self-deception will always have an essentially moral character, since the actions that constitute it are the appropriate target of moral judgments. Perhaps this seems unfair, since the consequences of self-deception are often beyond our control: you might think Mitt Romney could not reasonably have foreseen that his false beliefs would turn out to have such disastrous effects, and that this shows that his failing was epistemic, not moral. But our practices of holding one another responsible do not always track what is reasonable, as the copious literature on moral luck demonstrates. We are often held morally responsible for occurrences that are not, from a certain perspective, strictly up to us. Those practices can change over time, and thus the occasions on which we hold people responsible for being self-deceived can also change. But we can and do hold people like Mitt Romney for being self-deceived, and the explanation plausibly involves linking his intentional actions in sustaining self-deception with a theory of moral responsibility (as I will do in this chapter).

Furthermore, on a consequentialist theory of morality it isn't clear that the moral-epistemic distinction amounts to much. Engaging in certain evidence-gathering practices over others has certain morally relevant consequences, and moral norms reflect this. As I claim later on, there are moral norms specifically governing the activities whereby we acquire, revise, and sustain our beliefs. Sometimes those activities will have good consequences and other times not, and we should expect that for activities that have good consequences there will be norms in favor of engaging in those activities; for activities that have bad consequences there will often be norms against them. Thus it appears difficult to separate an activity that is purely epistemic from one that has moral implications – this is especially true on a consequentialist theory of morality, where *every activity* is morally significant.[18]

For anyone who is convinced that the moral-epistemic distinction is robust, I am quite happy to replace all talk of moral norms and requirements with talk of epistemic norms and requirements. Since presumably both kinds of norms can underlie reactive attitudes and thus support a practice of holding others responsible for self-deception, it does not much matter whether the responsibility bottoms out with distinctively moral requirements. The debate is irrelevant to my larger project of explaining what is special about paradigmatic self-deception.

4.2 Culpable Ignorance

When Le Roy parents rejected the diagnosis of conversion disorder in my example from Chapter III, they were perhaps ignorant of the fact that their beliefs in alternative explanations were empirically suspect. If so, can we blame them for that ignorance? Some cases of ignorance are clearly culpable, as Gideon Rosen (2003) notes: "[W]hen I walk down the sidewalk with my nose in a book, I am clearly acting recklessly. I am under an obligation to look out for other people when I'm out walking. If I recklessly shirk that obligation and wind up ignorant as a result, the ignorance is itself culpable, and in that case it's no excuse" (p. 63).

Being culpable for ignorance requires that the individual have acted *from* ignorance, which, following Alexander Guerrero (2007), means that in self-deception the ignorance must

[18] Virtue epistemologists may also reject the moral-epistemological distinction, as Olin and Doris (2014) note. For people like Ernest Sosa (1991), moral and epistemic virtues can be one and the same, since part of the "function" of a human being is to apprehend the truth and this can be done well or badly depending on one's intellectual virtues.

play an “important causal or explanatory role” in the phenomenon (p. 63). This is certainly true on the account of paradigmatic self-deception that I have given, since ignorance is what prevents the self-deceived individual who falsely believes that P from recognizing that not-P is actually true. But the trick now is to establish that the ignorance in paradigmatic self-deception is in fact culpable. Rosen talks about culpable ignorance of facts as often being associated with reckless or negligent behavior, so in that spirit I will treat the problem of culpable ignorance as a problem about negligence: a self-deceived person who is culpably ignorant *neglects* to make a reasonable effort to discover the fact of which he is ignorant. In the last chapter I claimed that paradigmatic cases of self-deception involve intentional ignorance. The task, then, is to uncover how being intentionally ignorant makes paradigmatically self-deceived individuals negligent.

4.2.1 The Possibility of Culpability

Before I talk about negligence, though, I should note that some philosophers who have claimed that we might *never* be culpable for ignorance. Michael Zimmerman (1997) distinguishes between direct and indirect freedom: “One is in indirect control of something, X, if and only if one is in control of it by way of being in control of something else, Y, of which X is a consequence…One is in direct control of something if and only if one is in control of it in some way that does not involve being in control of it by way of being in control of something else” (p. 415).

Invoking this distinction, Zimmerman claims that “no one can be in direct control of being ignorant” because “if one can avoid ignorance, one can do so only by way of doing something else by means of which ignorance is avoided” (pp. 415-16). [19] He also distinguishes

[19] In fact, if Zimmerman is right then sustaining ignorance is not an activity at all; ignorance is simply a state of being.

between direct and indirect forms of culpability, where indirect culpability is ultimately a product of direct culpability. As he writes, "[A]ll culpability can be traced to culpability that involves a lack of ignorance, that is, that involves a belief on the agent's part that he or she is doing something morally wrong" (p. 418).

For Zimmerman, then, we have only indirect control over ignorance and are thus at best indirectly culpable for it. Establishing the aptness of responsibility judgments in paradigmatic self-deception would ideally involve establishing direct culpability, but this seems unlikely on Zimmerman's account: the culpability of paradigmatic self-deception could not ultimately be rooted in intentional ignorance, as I believe, but rather in culpability for some other intentional activity that does not involve ignorance. Because I have claimed that paradigmatic self-deception has intentional ignorance as its central feature, merely indirect culpability for ignorance would mean that we can be only indirectly culpable for paradigmatic self-deception. The Le Roy parents were ignorant that conversion disorder was actually a good explanation for their children's illness, yet on Zimmerman's account they cannot be directly culpable for this even if their ignorance was caused – as I claim – by intentional activities.

Zimmerman's worry is related to what Manuel Vargas (2006) calls *tracing cases*:

> Oftentimes, we seem to trace responsibility for an action past the immediate structure of agency back to some earlier point in the agent's history. Drunk driving is one example. We hold someone responsible for the results of drunk driving not because of the kind of agent they are when they get behind the wheel, but rather, because of the kind of agent they were when they started to drink, knowing that drinking impairs judgment and coordination, and so on. That is, even if the agent does not satisfy the structural requirements for responsibility at the time of action, we can trace responsibility back to a prior state where the agent did satisfy the structural conditions in a way that generates responsibility for the later actions (pp. 356-7).

Zimmerman would claim that paradigmatic self-deception is like drunk driving: we can be culpable for paradigmatic self-deception only because we can trace the intentional ignorance of self-deception back to some culpable act that occurred before the self-deception began. But this claim is ambiguous. Zimmerman has in mind that this kind of culpability is indirect because the culpability is not attached to the ignorance itself but rather to the antecedent action. Strictly speaking, on this interpretation to say that I am culpable for my self-deception is to make a mistake, since the culpability attaches to an act that might have nothing particularly to do with self-deception.

I want to resist this conclusion. One way to do that would be to interpret Zimmerman and Vargas such that the culpability for self-deception could still be direct in that it attaches to the ignorance itself, while the *explanation* for the culpability refers to the antecedent action; that is, I am directly culpable for being self-deceived but *in virtue* of the antecedent action.[20] This strategy strikes me as little more than a semantic trick: if the antecedent action is what grounds the culpability, then what lets us say that the culpability doesn't attach to the antecedent action itself? Zimmerman would probably claim that it perverts the direct-indirect distinction, and I would tend to agree. If I am to claim that direct culpability for ignorance – for self-deception – is possible, I want to do it on Zimmerman's terms.

How to rescue direct culpability for ignorance? James Montmarquet (1999) argues question of culpability might not end up turning on control at all (p. 845):

> [A] certain quality of *openness* to truth- and value-related considerations is expected of persons and…this expectation is *fundamental,* at least in the following regard. The expectation is not derivative of or dependent upon one's (at the moment in question) judging such openness as appropriate (good, required, etc.)–just the opposite: it would include a requirement that one be open to the need to be open, and if one is not open to this, one may be blameworthy precisely for that failure.

[20] Thanks to John Doris (personal communication) for this point.

So even if it turns out that ignorance is not under direct control, Montmarquet thinks that agents can still be directly culpable for ignorance because there exists some general "expectation" of sensitivity to evidence. If there is some general societal norm that we be open to believing the truth of things instead of what we want to be true because it is most comfortable or expedient, then we will tend to hold each other morally responsible for any violations of that norm. The existence of this norm is an empirical question, but it seems relatively uncontroversial. Presumably most people think it important to have true beliefs about the world most of the time, especially when those beliefs concern matters we care about especially; whenever we see flagrant violations of that standard, then, we tend to become upset, even contemptuous. Those who put their trust in the Romney campaign's ability to win the 2012 election – or at least to compete competently – were upset when those leading it turned out to have disastrously wrong beliefs. If culpable ignorance is a matter of social norms, then individuals involved in paradigmatic cases of self-deception would plausibly be culpably ignorant.

But note that this does not provide any argument that the norm is *well-motivated*. The mere existence of a norm of openness does not demonstrate that individuals who follow it are responding to any objective features of the situations in which it is applied. One could be a Strawsonian and think that there is no such thing as direct control at all; a hard determinist about free will would fit this description, assuming direct control requires the ability to do otherwise. So we must be wary of concluding that paradigmatic self-deception is culpable merely based on a robust practice of treating it as such. There is still a difficult substantive question to resolve.

His other response is more compelling. Montmarquet charges that Zimmerman is too quick to give up on direct control over ignorance: "Direct control with respect to a modality – a way of doing something – does not imply indirect control with respect to that doing itself.

On the contrary, it would seem to imply direct – if by no means complete – control of the latter" (p. 844). He continues:

> [I]t should be clear that we can speak of an individual as having direct control over his degree of care and, to that extent, direct (albeit incomplete) control over his forming the belief that he is then forming. Thus, to say that he has only indirect control on the grounds that this belief is a consequence (in part) of his lack of due care is a non sequitur, akin to maintaining that I had only indirect control over my whistling because I would not have whistled that way had I exerted due care (p. 844)

This seems exactly right. I have claimed that paradigmatic self-deception is intentional in that it involves a pattern of intentional acts and omissions, and those acts and omissions are the means by which ignorance is sustained. That is, the intentional acts and omissions fully explain and constitute the ignorance involved in paradigmatic self-deception. How does this not amount to direct control over ignorance? Zimmerman's standard for directness is far too high. Even the truth of epistemic voluntarism, the theory that we can will arbitrary beliefs into existence merely by forming the intention to do so, would not be enough to make ignorance directly controlled on Zimmerman's account. Epistemic voluntarism is the most direct kind of control over ignorance imaginable, but since it still involves controlling ignorance *by way of* a capacity to manipulate beliefs it would seemingly still not meet Zimmerman's requirements.[21]

[21] This may be enough to show that the direct-indirect distinction is untenable. Similar worries apply to an argument from Neil Levy (2004) that self-deception is merely "indirectly" intentional. Levy correctly identifies the concern that self-deception could not possibly be intentional "under that description" (p. 299). That is, self-deceived individuals don't act on intentions with the content "DECEIVE MYSELF." Yet he immediately infers from this that self-deception can be at best "indirectly" intentional: "It is is an unforeseen (though sometimes welcome) consequence of cognitively evasive intentional activity" (p. 299). This explanation does not seem to apply to paradigmatic cases of self-deception as I have described them. Romney and the Le Roy parents certainly engaged in "cognitively evasive intentional activities, " as Levy calls them, but those activities – seeking out evidence that favored their false beliefs and omitting to find evidence that threatened them – but is unclear why this makes their self-deception indirectly intentional instead of just "intentional." On my view, such activities are the means by which which self-deceived individuals actively deceive themselves. Levy, like Zimmerman, seems to have a far too stringent standard for directness. I have claimed that, so long as the intentional content in self-deception is relevant to the activity, self-deception is just as intentional as as any other garden-variety intentional action; if this is right, the direct-indirect distinction is irrelevant and unhelpful.

Bratman and Harman would also surely disagree vehemently with Zimmerman as well in their examples from Chapter III: the runner directly controls the condition of his shoes and the sniper whether he alerts the enemy. So too do the Le Roy parents and the Romney campaign directly control their ignorance. Direct control over the modality of self-deception – over the acts and omissions that sustain ignorance – is enough to allow direct control over self-deception itself.

4.2.2 Responsibility Without Intention?

A second objection about the possibility of culpability arises in an important discussion of self-deception and responsibility from Neil Levy (2004). Levy's overall view is that self-deception amounts to a simple mistake, and that "although people are sometimes responsible…for their mistakes, there is no presumption of such responsibility" (p. 295). He arrives at this conclusion after arguing against both the two-beliefs requirement, according to which self-deception involves holding contradictory beliefs, and the intentionality requirement, which says that self-deception is "intentionally produced" (p. 295). As Levy rightly notes, these two requirements distinguish self-deception from mere mistaken belief, so responsibility for self-deception seems to depend on at least one of them being true. I defended both at length in the last chapter and won't reiterate my arguments here. The challenge is that Levy believes that culpability for self-deception might obtain even if it involves neither holding contradictory beliefs nor intentionally deceiving oneself. Success here would pose a serious problem for the project I have been advancing. If we can get moral responsibility without intentionality, then there is no puzzle about the way in which paradigmatic self-deception can be culpable; intention would end up being superfluous, and the last chapter would be a waste of argument.

Levy claims that responsibility for self-deception will follow just when (1) “the subject matter of the belief is important (whether morally or in some other manner)” and (2) “we are in some doubt about [the belief’s] truth” (p. 305). Curiously, he seems to think that this doesn’t require much independent argument, saying only that (1) and (2) also describe when we “remind ourselves to consider both sides of a question” (p. 305). The intuition seems to be that we can be culpable for self-deception only if the false belief arises out of some kind of deliberative process – a process designed to eliminate doubt – about some morally weighty or otherwise important question.

Yet this doesn’t seem to match at all the theory of paradigmatic self-deception that I have discussed. We can be self-deceived about all kinds of things, important or not, since the false belief is produced via motivated reasoning processes that are driven by desires, which don’t track any objective notion of importance. So paradigmatic cases will only occasionally satisfy the importance condition. But why have such a condition at all? If the theory of paradigmatic self-deception that I have given is right, then responsibility will be tied to the sharpness (of the false belief), diachronicity, and effort involved in a particular case. To the extent that responsibility comes from what a person *does* to deceive herself, incidental features such as importance are irrelevant; the importance condition is hard to take seriously.

Of (2), Levy rightly notes it seems implausible that self-deception will involve entertaining doubts, saying that “[c]oncurrent doubts are ruled out almost by definition: effective self-deception seems to *preclude* the concurrent satisfaction of (2)” (p. 307). He goes on to use this point to criticize Annette Barnes (1997), who makes doubt a central feature of her theory of self-deception. Since Levy gives no independent argument for (2), it seems mostly useful as a tool for criticizing Barnes: he asserts without evidence that (2) is required for responsibility and explains how Barnes’ proposal, which places doubt at the center of self-deception, ultimately fails. But the truth of (2) is hardly obvious. If (1) and (2) were really conditions on responsibility for self-deception then it would be no surprise that responsibility ends up being difficult or even impossible. I argue that Levy suffers from a failure of imagination: as I am about to show, we can get to responsibility without either of his conditions.

4.3 Negligence and Duty of Care

Having responded to these objections, I return to the original question of negligence. The concept of negligence is most salient in the literature on criminal and tort law. When I cause harm to you, should I be found liable? Under one legal standard, *strict liability*, the answer is typically yes: regardless of the context or any extenuating circumstances, I owe you restitution. But on another, the *negligence* standard, I owe restitution only if in causing you harm I acted negligently. But how do we cash out negligence? Following Joseph Raz (2010), I conceive of negligence as a violation of a *duty of care*, "the standard of care people have to observe in their dealings with each other" (p. 6). Violations of this duty entail moral responsibility:

> Negligent conduct is not careless conduct; it is careless conduct for which one is responsible, and where care was due. It is a conceptual truth that negligent conduct is a *prima facie* wrong for which one is responsible. One can rebut allegations of negligence by establishing that one is not responsible for the conduct, or by establishing that even though the conduct was careless, one had no reason to be careful (p. 9).

So it follows that to show that a self-deceived agent is negligent – and thus morally responsible – I must show that the agent violated a duty of care.

Which duty, exactly? I need to define a particular duty that is relevant to paradigmatic self-deception. I propose that paradigmatically self-deceived individuals violate a duty *to take relevant evidence seriously*, even – especially – when doing so is uncomfortable. If, as I proposed in the last chapter, paradigmatic self-deception is sustained via certain attentional capacities, then taking evidence seriously would amount to attending to it in the right way and to the right extent. If we instead wanted to explain taking evidence seriously as some kind of reflective notion, then it would amount to giving evidence the proper weight in reflective deliberation. (I don't think reflection is important for understanding self-deception, but the duty of care in self-deception can be neutral as to which theory of self-deception is correct.)

I won't attempt to give necessary and sufficient conditions for what it is to take evidence seriously in general. But our discussion of mindfulness from the last chapter can help us understand what it is to take evidence seriously in the context of self-deception. I have claimed that sustaining self-deception plausibly involves using the same kinds of attentional skills that mindfulness practitioners do: self-deceived individuals fluently direct their attention to focus on the false belief and the evidence for it, making it easier to shut out anything that might threaten it. This, I claim, is a failure to take evidence seriously.

So whenever there is a case of paradigmatic self-deception there is also a case in which someone failed to take evidence seriously, either by ignoring it or by giving it insufficient weight. The converse is also true: whenever similarly situated agents grapple with the same evidence and avoid self-deception, then they took the evidence seriously.[22] But what is it to do that? Recall that one skill in mindfulness is *sustained attention*: the practitioner focuses for an extended period on the present moment and acknowledges all of the thoughts in the present stream of consciousness. Taking evidence seriously might involve sustaining attention on all of the thoughts – the bits of *evidence* – that are relevant to the question at hand. If *not* taking evidence seriously involves focusing on some evidence to the exclusion of everything else, then taking evidence seriously would require, at a minimum, attending to the body

[22] That is, a person in a similar epistemic position to the self-deceived person.

of evidence as a whole. This is just to say that taking evidence seriously involves *not* making the omissions that are characteristic of paradigmatic self-deception.

Consider again Mitt Romney and his campaign staff, who intentionally avoided considering evidence that threatened their false belief in their polling model and in so doing were intentionally ignorant. The same was true of the Le Roy parents, who were intentionally ignorant that their children were quite plausibly suffering from a conversion disorder coupled with mass psychogenic illness. In cases like these, which I believe are representative of paradigmatic cases in general, the violation of a duty to take evidence seriously is most obvious because of the *contrast* between those who were self-deceived and those who were not. Many political pundits predicted an Obama victory in 2012 and trusted the mainstream polling that guided that prediction. Some parents in Le Roy believed the official diagnosis and took their children to a neurologist, who treated them with cognitive-behavioral therapy. The duty to take evidence seriously exists for everyone, but only some are successful in following it. And the fact that in these cases some people – a great many, in the case of the Romney campaign – were able to fulfill their duty suggests that doing so is in general possible. Absent evidence to the contrary, we should assume that it is psychologically possible for *anyone* to take evidence seriously and thereby avoid paradigmatic self-deception.

The duty to take evidence seriously explains and justifies our general practice of holding people responsible. You might think, though, that the Romney and Le Roy cases aren't especially good evidence for the existence of a general duty. Negative reactions to Romney's loss could be explained as broad Republican disappointment with an unfavorable electoral outcome, not as specific outrage at the campaign's choice of a bad polling model. And many people might be reluctant to blame the Le Roy parents who, though they were self-deceived, genuinely believed that they were acting in their children's best interests.

So let's consider a different case. The Transportation Security Administration (TSA) is a branch of the U.S. Department of Homeland Security charged with ensuring the safety of the traveling public, most notably by running security checkpoints at commercial airports around the country. Since 2007, the TSA has deployed Behavioral Detection Officers (BDOs) at checkpoints to watch for passengers displaying certain "suspicious" behaviors and subject those passengers to additional screening. The program has spent over $1 billion since its introduction, and the BDOs' training is based on behavioral theories developed by the psychologist Paul Ekman, who claims that "microexpressions" reveal underlying emotions in ways that BDOs can recognize and exploit.[23] A 2008 post on the TSA's official blog explained how BDOs operate (Burns, p. 1):

> Behavior analysis is based on the fear of being discovered. People who are trying to get away with something display signs of stress through involuntary physical and physiological behaviors. Whether someone's trying to sneak through that excellent stone ground mustard they bought on vacation, a knife, or a bomb, behavior detection officers like me are trained to spot certain suspicious behaviors out of the crowd.

There's only one problem: the program doesn't work. Ekman's ideas about the connection between emotions and behavior have not survived experimental testing, as the Government Accountability Office (GAO) warned in a 2013 report to Congress on the program's effectiveness:

> GAO reviewed four meta-analyses (reviews that analyze other studies and synthesize their findings) that included over 400 studies from the past 60 years and found that the human ability to accurately identify deceptive behavior based on behavioral indicators is the same as or slightly better than chance. Further, the Department of Homeland Security's (DHS) April 2011 study conducted to validate SPOT's behavioral indicators did not demonstrate their effectiveness because of study limitations, including the use of unreliable data (p. 1).

[23] Ekman and his work were the basis for the 2009-2011 FOX drama series *Lie To Me*.

The GAO concluded that "[u]ntil TSA can provide scientifically validated evidence demonstrating that behavioral indicators can be used to identify passengers who may pose a threat to aviation security, the agency risks funding activities that have not been determined to be effective" (USA 2013, p. 1).

Despite evidence of the program's ineffectiveness, TSA director John S. Pistole has ardently defended it to Congress and the public. In 2013 Congressional testimony he rebuked the GAO in asserting that "BDOs are trained to identify behavior cues that have been shown through research, science, and decades of domestic and international law enforcement experience to be reliable indicators and predictors of anomalous or suspicious behavior" (Pistole 2013). Pistole is also fond of anecdotes, as in a 2012 *USA Today* op-ed: "In the past month alone, TSA officers trained to pick up on behavioral cues saved a beaten and kidnapped woman from her kidnappers in Miami and came to the rescue of a man having a heart attack in Boston" (Pistole 2012, p. 1).

Outrage has intensified as of late. In the *New York Times*, John Tierney conducted damning interviews with psychologists who chalked up Pistole's enthusiasm to a naive belief in our ability to "read liars' minds by watching their bodies" (Tierney 2014, p. 1). "'There's an illusion of insight that comes from looking at a person's body,' says Nicholas Epley, a professor of behavioral science at the University of Chicago. 'Body language speaks to us, but only in whispers'" (Tierney 2014, p. 1). The independent blog *TSA News* was blunter: "The TSA has just spent over a billion of our tax dollars only to find out that Pinocchio's nose is make-believe" (Tornello 2014, p. 1).

From the facts of the case, Pistole is obviously self-deceived about the program's effectiveness. He focuses exclusively on the program's "successes" while waving away the decisive evidence that its theoretical foundations are bogus.[24] In this case, the negative reactions from Congress, the press, and the public seem directed at the self-deception itself. The federal government loses many billions of dollars annually to waste, fraud, and abuse, but here the money is being wasted entirely because Pistole has failed in his duty to take evidence seriously in the course of executing his job. Here, the problem with a costly, mismanaged government program can be reduced to something simple: self-deception.

I claim that this case decisively demonstrates the existence of a duty for Pistole to take evidence seriously. Part of the TSA director's job is to make sound, evidence-based decisions about how to spend the agency's limited budget for the greatest benefit to the American people. The BDO program reveals that Pistole has fallen down on the job, both by making a bad initial judgment in authorizing the program and also by continuing to defend it even in the face of strong evidence that the theory behind it is flawed. So Pistole violated – and continues to violate – one of the duties of his office, and the press and the public are justifiably outraged.

From this particular case, we can infer that the duty to take evidence seriously is widespread. While being TSA director carries with it the special burden of making decisions that affect the lives of millions of people, the principles of good decision making are universal. Pistole's error was in a sense specific to his job: he is self-deceived about the effectiveness of a particular TSA program. But in another sense his error is of exactly the same kind that countless other people make every day in forming judgments in their personal and professional lives. For Pistole the stakes were considerably higher, and so he is likely deserving of some special outrage for the magnitude of the bad consequences. But the bulk of the outrage

[24] The anecdotes Pistole cites as demonstrating the program's effectiveness have many explanations that do not validate Ekman's theories. Simply having officers watching for suspicious behavior instead of screening passengers at random likely has significant benefits, since nervous people behave in characteristic ways that have nothing to do with "microexpressions." So perhaps not all of the $1 billion has been wasted, only the big chunk allocated for the special training.

attaches to his decision process itself – how could he have been so careless? And thus we can plausibly conclude that outrage was appropriate also for Romney's bad judgment, and for the Le Roy parents', and for the judgment of anyone else who is self-deceived and lacks an excuse: how could *they* all have been so careless? That is, in our society there is a general duty to take evidence seriously. Violations of the duty are not always punished, both because many cases of self-deception go unnoticed and because self-deceived people may also be guilty of other, more serious crimes.[25] But the duty exists nonetheless.

4.4 How the Self-Deceived Are Negligent

Now assuming that the duty to take evidence seriously exists, It will help to get a little clearer on the sense in which paradigmatically self-deceived individuals violate it and can reasonably be expected not to do so. Some philosophers who discuss negligence and responsibility defend what Steven Sverdlik (1993) calls the Aristotelian Strategy, "look[ing] for an act in the past that produced the ignorance or error operative at the time of the norm violation" (p. 139). The strategy is broadly informed by the *Nicomachean Ethics*, in which Aristotle writes that we can hold someone responsible for ignorance when he was the cause of his ignorance: "Indeed, we punish a man for his very ignorance, if he is thought responsible for the ignorance, as when penalties are doubled in the case of drunkenness; for the moving principle is in the man himself, since he had the power of not getting drunk and his getting drunk was the cause of his ignorance" (III, 5).

[25] Much of the unhappiness after Romney's loss may be attributed to partisan disappointment. But I claim that some of it was – or at least *should have been* – targeted specifically at their violation of the duty to take evidence seriously.

Holly Smith (1983) and Michael Zimmerman (1986) have both defended versions of this view.[26] Zimmerman thinks that negligent behavior involves an unjustifiable case of failing to consider a risk at the moment of action coupled with consideration of that risk at some earlier time. He asks us to imagine Bert the bricklayer, who has a habit "of tossing defective bricks over his shoulder" that one day leads to the death of a pedestrian when a brick that Bert has tossed from a skyscraper hits him on the head (p. 199). Obviously Bert didn't consciously entertain the risk of hurting someone just before tossing the brick; if he had, the case would shade into recklessness instead of negligence. But Zimmerman claims that it is only a case of negligence if Bert had, at some earlier time, considered that "he might engage in such an activity and thereby cause damage or injury" (p. 200).[27] Smith (1983) offers a helpful encapsulation of the view, which decomposes a case of negligence into a pair of acts: "an initial act, in which the agent fails to improve (or positively impairs) his cognitive position, and a subsequent act in which he does wrong because of his resulting ignorance" (p. 547). The initial act he calls the *benighting act*.

The Aristotelian Strategy doesn't seem to describe paradigmatic self-deception well. Recall that Neil Levy thinks responsibility for self-deception requires that the self-deceived person be in doubt about the truth of false belief. Adopting the Aristotelian Strategy for negligence would require that in paradigmatic self-deception the individual considered some risk that she might fail to take evidence seriously about some question or other and then unwittingly failed to take evidence seriously when that question arose at some later time. This

[26] Acknowledgment is due to Steven Sverdlik (1993) for bringing these papers to my attention.

[27] Zimmerman's view here is of a piece with his views on culpable ignorance we saw earlier. Presumably he favors the Aristotelian Strategy because the benighting act corresponds to the act that produces ignorance and thus accrues culpability; recall he thinks ignorance is indirectly culpable because it is controlled via some other means. He cannot endorse Sverdlik's theory of pure negligence because doing so would reduce negligence to a single act and ignorance could be directly culpable.

seems very much like Levy's doubt requirement, since the most plausible way of cashing out the benighting act is as a case in which the self-deceived person entertains and perhaps suppresses doubts about how she will respond to certain evidence relevant to the false belief that she ends up adopting and sustaining in the activity of paradigmatic self-deception.[28] While a paradigmatically self-deceived agent *might* have performed such a benighting act, there is nothing in the cases I have discussed that *requires* that the agent have done so. That is, we have no special reason to think that the Aristotelian Strategy for negligence captures what goes on in paradigmatic self-deception.[29]

You might also think that the Aristotelian Strategy threatens to be self-undermining in self-deception. If the benighting act involves entertaining and suppressing doubt about the possibility of being self-deceived, or about the possibility of coming to hold the relevant false belief, then any memory of the act that resurfaces would seemingly make it impossible to be ignorant in the way that constitutes self-deception. Consider the Le Roy parents: if we suppose that they were negligent because of some benighting act in which they suppressed doubts about their susceptibility to believing spurious scientific claims or failed to improve their understanding of what constitutes good scientific evidence, then remembering that they had done so would make it difficult – perhaps impossible – to sustain their false beliefs in the long run. So for the self-deception to be successful the Le Roy parents must direct their attention away from not only the evidence against their false belief but also any evidence of their benighting act, which would also threaten the false belief. We only need to suppose that self-

[28] Smith (1983) also emphasizes that the benighting act can often be an omission, in which case the benighting act in paradigmatic self-deception could be an omission to consider the risk.

[29] The Aristotelian Strategy also seems to make paradigmatic self-deception into a tracing case of the kind Manuel Vargas (2006) describes. Because there exists some benighting act, we can trace culpability back to that act and thus the culpability really accrues to that act and not to the activity of self-deception itself.

deceived individuals are doing all this additional work if negligence requires a benighting act. But there is an alternative theory of negligence.

I want instead to characterize paradigmatic self-deception as cases of what Sverdlik (1993) calls *pure* negligence, or negligence for which there is moral responsibility despite the absence of any benighting act. He motivates his rejection of the Aristotelian Strategy by identifying a fallacious inference, F, that may be lurking behind that view: "I can deliberately act to intervene into my cognitive processes, and thereby eliminate some of my ignorance or false beliefs; therefore, the ignorance or errors that would have existed had I not intervened were the product of a deliberate act" (p. 142). Reading "deliberately" as "intentional," F is clearly illegitimate, since the mere fact that I can intentionally act to, say, reduce the chances that I will be self-deceived doesn't show that not doing so makes my self-deception intentional. Claiming otherwise amounts to an illicit attempt to smuggle intentionality into the account of negligence. Given that F is associated with Aristotelian Strategy of explaining negligence, rejecting F gives us decisive reason to prefer a pure negligence account over the Aristotelian Strategy; my endorsement of pure negligence is not *ad hoc*.

Levy (2004) might himself have fallen into the trap of believing F. He reads Mele as "insist[ing] that self-deceivers are typically responsible," even on a deflationary account like Mele's own, because "the kind of biasing mechanisms usually at work [in self-deception] are to some extent under our control….[I]f we know about the confirmation bias, we can…neutralize or at least minimize its effect on our thinking" (pp. 304-5). But Mele makes no such claim, at least not explicitly. In the passage Levy cites, Mele (2001) merely notes several cases in which knowledge of biases can reduce the effects of those biases and says that this kind of control is "a resource for combating self-deception" (p. 103). Mele says nothing about whether such control is sufficient for moral responsibility – whether being affected by those biases is the result of an intentional act. This is probably because being able to minimize the effects of a particular bias upon notification of how exactly it is affecting my reasoning in a particular case doesn't provide much evidence for a *general capacity* to minimize the effect of that bias. Controlling bias to any significant degree might well depend on having specific information that is typically unavailable; after all, there is no evidence that cognitive psychologists are less susceptible to bias in their daily lives than everyone else, despite studying reasoning errors for a living.

In interpreting Mele as he does, Levy seems to rely on Sverdlik's inference F to wrongly conclude that deliberate control over a particular intervention into bias entails that *not* intervening is also deliberate, intentional, and culpable. We cannot make the same mistake in analyzing the culpability of paradigmatic self-deception. In order to validate the Aristotelian Strategy, the omission to act to reduce ignorance or error in self-deception would itself necessarily be intentional in order to guarantee that the benighting act is culpable. But there is no good reason to think that an intentional omission to consider the risks of self-deception precedes every instance of self-deception.

Pure negligence, then, is a theory that avoids relying on F. Instead of supposing that negligence decomposes into a pair of acts, the earlier of which is the true source of culpability, Sverdlik argues that the "locus of blame" in negligence falls on the fact that the individual failed to intervene "into the processes that were leading to a situation of an unwitting norm violation," where such an intervention was *possible* for her (p. 142). The individual could have intervened in a way that would have prevented the norm violation – violating the norm of taking evidence seriously, in our case – but she didn't, and that's why her negligence is culpable; this differs from the Aristotelian Strategy since the norm violation need not itself be intentional for culpability to follow.

Sverdlik thinks a successful intervention would involve “moral reflection,” in which “the conscientious agent…surveys the situation and the appropriate questions or beliefs ‘present themselves’” (p. 142). But moral reflection doesn’t sound promising in the context of the theory of paradigmatic self-deception I have described, primarily because I think self-deception is sustained through directed exercise of attention rather than via any reflective capacities.[30] Instead of moral reflection, the relevant intervention in this case would be directing attention in ways that make it harder – even better, impossible – to be self-deceived about a particular matter. Because paradigmatic self-deception is a temporally extended activity, I also need not require that the intervention was possible prior to the beginning of the self-deception. Sverdlik implicitly assumes that the norm violation in a case of negligence will be synchronic – a single, isolated incident – and so the only possible intervention that could avert a norm violation would be one that took place prior to the violation. But since paradigmatic self-deception is an ongoing activity, we should instead focus on what goes on *during* it. That is, paradigmatic self-deception is a case of pure negligence because the self-deceived individual could have intervened to prevent the norm violation – the violation of the duty to take evidence seriously – from *continuing*.

How do we know that such an intervention is possible? In the last chapter I argued that paradigmatic self-deception is intentional and sustained via various directed exercises of attention. Since attention is subject to intentional control, I claim that at any stage it is *possible* to intentionally direct attention in ways that would collapse the self-deception. This is not to say that *not* directing attention in that way is at every moment an intentional act – saying

[30] As I argued in chapters II and III, we can fully explain paradigmatic self-deception without ever bringing in reflection. The attentional capacities I’ve described as central to self-deception don’t require reflection to direct them, and since we have direct, intentional control over them it’s unclear what work reflection is supposed to do.

otherwise is to make Sverdlik's fallacious inference F – but on the theory of pure negligence the *possibility* of intervention is all that is required for negligence. If the arguments of the last chapter are successful, then the possibility is certainly present. And failing to intervene in this way violates the duty to take evidence seriously and thus makes self-deception negligent and thus culpable.

4.5 Conclusion: Negligence, Mistakes, and Recklessness

Should the conclusion that we have moral responsibility for the negligence of self-deception seem suspicious, I now want to place negligence along a continuum of acts for which we have more and less moral responsibility. Negligence is morally worse than making a simple mistake but much less bad than recklessness. Suppose a case of paradigmatic self-deception involved not negligence but a mistake: the Romney campaign used a bad polling model not because they were intentionally, negligently ignorant that it was bad but because they made a mathematical error in calculating how likely their polling model was to match the actual November 2012 turnout. Perhaps occasionally mistakes like these are culpable if they are produced by motivated reasoning, but often it's just a simple mistake: human reasoning is fallible and sometimes malfunctions, and when this happens the conditions for moral responsibility are not satisfied. Even if through some incredibly unfortunate string of errors the polling model was used all the way up until the election and so directly led to Romney's loss, this would in no way approach the badness of being truly negligent.

Recklessness is much worse than negligence. If the Romney campaign chose a bad polling model with *full knowledge* that it was likely to fail in predicting actual turnout, then their culpability would be far higher than in a case of negligence. The overall point is that self-deception is hardly the worst crime they could have committed. We can and should blame people for morally culpable self-deception, but it exists on a spectrum of greater and lesser culpability. What is important is that paradigmatic self-deception is generally culpable, absent any excusing conditions. This is not obvious even given the arguments of the last chapter, and on many accounts of self-deception it is impossible.

On a more hopeful note, if I am right that it is always possible to intervene in the processes that give rise to self-deception, then paradigmatic self-deception is in general *curable*. I cannot help but reason in certain ways that give rise to false beliefs motivated by my desires, but I do have the power to prevent those false beliefs from taking over my life. Sometimes that will be desirable, as in our relationships with those closest to us it can be beneficial – perhaps even morally required – to sustain false beliefs that help the relationship flourish. Other times it is certainly very bad. Not being self-deceived may have made all the difference to the Romney campaign's chances in 2012, and it would certainly have helped the sick children in Le Roy get the necessary treatment sooner. Examples like theirs show us how powerful the combination of false beliefs and negligent behavior can be.

Epilogue
Your Problem and Mine

I have tried in this dissertation to shed some light on a psychological condition that, despite much prior discussion, was still frustratingly under-explained. I also recognize that this dissertation has been a bit of a downer. Self-deception is terribly widespread, and the paradigmatic variety is responsible for lots of harmful, painful, and just plain bad situations that have affected the lives of you and everyone you know.[1] Even more distressingly, the pervasiveness of paradigmatic self-deception threatens our self-conception as rational creatures. In ordinary self-deception, rationality may seem secure: a false belief arises from motivated reasoning, but the conscious mind, the part that's "really" you, isn't involved in producing that belief. So we can say that false belief might not be one that you would endorse. Human beings are saddled with unconscious biases as a result of our evolutionary history, but those biases are separate from the rational processes that otherwise guide our reasoning. Ordinary self-deception and rationality can thus co-exist.

But *paradigmatic* self-deception is different. If what I have argued is correct, then paradigmatically self-deceived individuals intentionally self-deceive by sustaining false beliefs despite evidence that ought to convince them otherwise. The desire that produced the false belief continues to drive their intentional behavior, which, as I argued in Chapter IV, makes them apt for judgments of moral responsibility. Insofar as rationality requires taking relevant evidence seriously and not treating it ways that are convenient for our desires, paradigmatic

[1] I duly note cases of paradigmatic self-deception that have positive consequences. In such cases, the prescriptions in this epilogue are probably not worth following.

self-deception is incompatible with rationality – in our susceptibility to paradigmatic self-deception, we reveal ourselves to be irrational.[2]

Is their any room left for optimism? It all depends on how deep the tendency to engage in paradigmatic self-deception really goes, a question that I am not equipped to fully answer. More specifically, it depends on whether self-deception is ineradicable. If it is possible to make genuine progress in reducing an individual's susceptibility to self-deception, then there is hope that over time a person who was once horribly self-deceived can escape his self-deception and not fall prey to it in the future. This is an empirical question, but a good start would be to seek greater public awareness of the possibility and dangers of self-deception and to point to cases like the ones I have discussed. While the cognitive biases literature has filtered into the public consciousness to the point that it is common to accuse people we disagree with of "projecting" or being victims of "cognitive dissonance," my arguments in this dissertation demonstrate that an understanding of biases alone is insufficient to grasp the ways in which human reasoning can go awry in self-deception. Paradigmatic self-deception merely begins with bias and is then sustained by intentional action. Fairly and completely characterizing the whole phenomenon of self-deception for a general audience thus requires careful explanation of how bias and intentional action can intermingle in unintuitive and unsettling ways. Greater public awareness of the mind-boggling complexity of paradigmatic self-deception can help jumpstart the process of reducing self-deception's overall incidence, assuming that such a reduction is possible. Once we know more about how we are flawed, we can cultivate solutions that let us avoid those flaws and even eliminate them.

In the meantime, I want to close with two practical conclusions we can draw from the foregoing discussion. First, avoiding homogeneous ideological groups may be an excellent

[2] The definition of rationality is, of course, up for grabs.

strategy for preventing some potential cases of paradigmatic self-deception. In many of the cases I have pointed to – the NASA engineers, Romney, the Le Roy parents, TSA Director Pistole – the self-deceivers were abetted in self-deception by being surrounded with like-minded people who did not challenge their false beliefs. Since paradigmatic self-deception is, I have argued, a matter of finding and omitting to find relevant evidence, the activity of self-deception is made considerably easier by being in an environment where threatening evidence is scarce. Romney surely would have had a much tougher time believing his polling model was correct if he hadn't been surrounded by a political apparatus – not to mention a huge swath of the American electorate – that was cheering for him to win and not at all interested in presenting their candidate with evidence that his fundamental strategic assumptions were wrong. Self-deception is easier in a like-minded group, so to avoid it we should try to spend serious time – maintaining an uncynical and open-minded attitude – around people who disagree with us on a wide range of political, personal, and professional issues.

Second, the phenomenon of paradigmatic self-deception demands *modesty*. Perhaps the most disturbing feature of self-deception is that we lack first-person awareness that we are self-deceived. Given that, it behooves us all to be mindful that our deepest convictions can sometimes be grounded in nothing like the sober, all-things-considered deliberative conclusions that we like to think form the bedrock of the beliefs that most matter to us. Introspection and subjective certainty is no guide at all to the likelihood that we are self-deceived. Modesty thus requires that when we argue with others we stop to consider – far more carefully than we might want to – that the passion and righteousness we feel when we defend what we feel to be right and true just might be founded in a big mistake. When it comes to paradigmatic self-deception, hesitation is a virtue.

If we learn nothing else, let it be the following: it is both presumptuous and pernicious to look at paradigmatic self-deception and believe that it is something that only happens to other people, to weak-willed or stupid people. Be assured that people at least as smart and strong of will as you have fallen victim to it, and there is no good reason to think that you too are not a victim now or will be in the future. Self-deception is both your problem and mine.

References

Achtziger, A., Gollwitzer, P. M., & Sheeran, P. (2008). Implementation intentions and shielding goal striving from unwanted thoughts and feelings. *Personality and Social Psychology Bulletin, 34*(3), 381-393.

Aristotle (1998). *Nicomachean ethics* (M. Pakaluk, Trans.). Oxford: Clarendon Press.

Audi, R. (1982). Self-deception, action, and will. *Erkenntnis, 18*(2), 133-158.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91*(4), 612-625.

Bargh, J. A. (2005). Bypassing the will: Towards demystifying the nonconscious control of social behavior. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 37-58). Oxford: Oxford University Press.

Barnes, A. (1997). *Seeing through self-deception*. Cambridge: Cambridge University Press.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., . . . Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice, 11*(3), 230-241.

Block, N. (2010). Attention and mental paint. *Philosophical Issues, 20*(1), 23-63.

Boyd, R. N. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on moral realism* (pp. 181-228). Ithaca, NY: Cornell University Press.

Bratman, M. (2000). Reflection, planning, and temporally extended agency. *The Philosophical Review*, *109*(1), 35-61.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review, 93*(3), 375-405.

Burns, B. (2008, February 29). The truth behind the title: Behavior detection officer [Web log post]. Retrieved from http://blog.tsa.gov/2008/02/truth-behind-title-behavior-detection.html

Clarke, R. (2010). Intentional omissions. *Noûs, 44*(1), 158-177.

Coffman, S. J., Dimidjian, S., & Baer, R. A. (2006). Mindfulness-based cognitive therapy for prevention of depressive relapse. In R. A. Baer (Ed.), *Mindfulness-based treatment* *approaches: Clinician's guide to evidence base and applications* (pp. 3-30). London: Academic Press.

Collins, R. L., Taylor, S. E., Wood, J. V., & Thompson, S. C. (1988). The vividness effect: Elusive or illusory? *Journal of Experimental Social Psychology*, *24*, 1-18.

Crawford, J. (2012, November 8). Adviser: Romney "shellshocked" by loss. Retrieved from http://www.cbsnews.com/news/adviser-romney-shellshocked-by-loss/

Davidson, D. (1982). Two paradoxes of irrationality. In R. Wollheim & J. Hopkins (Eds.) *Philosophical essays on Freud* (pp. 289-305). Cambridge: Cambridge University Press.

Davidson, D. (1987). Deception and division. In J. Elster (Ed.), *The multiple self* (pp. 79-92). Cambridge: Cambridge University Press.

Davidson, D. (1998). Who is fooled? In J. Dupuy (Ed.), *Self-deception and paradoxes of rationality* (pp. 1-18). Stanford: CSLI Publications.

Demos, R. (1960). Lying to oneself. *The Journal of Philosophy,* 57(18), 588-595.

Dennett, D. C. (1969). *Content and consciousness*. London: Routledge & K. Paul.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Dominus, S. (2012, March 7). What happened to the girls in Le Roy. *The New York Times Magazine*. Retrieved from http://www.nytimes.com/2012/03/11/magazine/teenage-girls-twitching-le-roy.html

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*(1), E29081.

Dunning, D. (2007). Self-image motives and consumer behavior: How sacrosanct self-beliefs sway preferences in the marketplace. *Journal of Consumer Psychology*, *17*(4), 237-249.

Fingarette, H. (1969). *Self-deception*. London: Routledge & K. Paul.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1030-1044.

Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives, 21*(1), 231-258.

Glaeser, E. L. and Sunstein, C. R. (2013). Why does balanced news produce unbalanced views? NBER Working Paper No. w18975. Available at SSRN: http://ssrn.com/abstract=2254227

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54*(7), 493-503.

Guerrero, A. A. (2007). Don't know, don't kill: Moral ignorance, culpability, and caution. *Philosophical Studies, 136*(1), 59-97.

Gulley, N. (2012, February 04). Parents urge more tests as twitches spread at New York school. Retrieved from http://www.reuters.com/article/2012/02/04/us-students-tics-newyork-idUSTRE8130S020120204

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology, 37*(2), 147-169.

Hamlyn, D. W. (1971). *The theory of knowledge*. London: Macmillan.

Harman, G. (1976). Practical reasoning. *Review of Metaphysics*, *29*, 431–463.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE, 8*(8), E72467.

Heil, J. (1989). Minds divided. *Mind, XCVIII*(392), 571-583.

Heil, J. (2003). *From an ontological point of view*. Oxford: Clarendon Press.

Hornsby, J. (2000). Personal and sub-personal: A defence of Dennett's early distinction. *Philosophical Explorations, 3*(1), 6-24.

Jack, A. I., & Robbins, P. (2004). The illusory triumph of machine over mind: Wegner's eliminativism and the real promise of psychology. *Behavioral and Brain Sciences, 27*(05).

James, W. (1890). *The principles of psychology*. New York: H. Holt and Company.

Johnson, G. (2012, November 8). Mitt Romney planned Boston Harbor fireworks show that was scotched by election loss. Retrieved from http://www.boston.com/politicalintelligence/2012/11/08/mitt-romney-planned-boston-harbor-fireworks-show-that-was-scotched-election-loss/qmgtVKPq4zNnDyb9FbLWeJ/story.html

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies, 68*(3), 221-263.

Keller, S. (2007). *The limits of loyalty*. Cambridge: Cambridge University Press.

Kihlstrom, J. F. (2008). The automaticity juggernaut–or, are we automatons after all. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 155-180). Oxford: Oxford University Press.

King-Farlow, J. (1963). Self-deceivers and Sartrian seducers. *Analysis, 23*, 131–36.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge: Cambridge University Press.

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford: Oxford University Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-498.

Kunda, Z., & Sinclair, L. (1999). Motivated reasoning with stereotypes: Activation, application, and inhibition. *Psychological Inquiry*, *10*(1), 12-22.

Lazar, A. (1998). Division and deception: Davidson on being self-deceived. In J. Dupuy (Ed.), *Self-deception and paradoxes of rationality* (pp. 19-36). Stanford: CSLI Publications.

Levy, N. (2004). Self-deception and moral responsibility. *Ratio, 17*(3), 294-311.

Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences, 12*(4), 163-169.

McLaughlin, B. P. (1988). Exploring the possibility of self-deception. In B. P. McLaughlin & A. O. Rorty (Eds.), *Perspectives on self-deception* (pp. 29-62). London: University of California Press.

Mele, A. R. (1987). Recent work on self-deception. *American Philosophical Quarterly*, *24*(1), 1-17.

Mele, A. R. (1992). Recent work on intentional action. *American Philosophical Quarterly*, *29*(3), 199-217.

Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences, 20*, 91-136.

Mele, A. R. (2001). *Self-deception unmasked*. Princeton: Princeton University Press.

Mele, A. R., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, *82*(3), 265-287.

Montmarquet, J. (1999). Zimmerman on culpable ignorance. *Ethics, 109*(4), 842-845.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science, 12*(1), 9-17.

Murray, S. L. (1999). The quest for conviction: Motivated cognition in romantic relationships. *Psychological Inquiry*, *10*(1), 23-34.

Neisser, U. (1979). The control of information pickup in selective looking. In A. D. Pick (Ed.), *Perception and its development: A tribute to Eleanor J. Gibson* (pp. 201-219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Olin, L., & Doris, J. M. (2014). Vicious minds: Virtue epistemology, cognition, and skepticism. *Philosophical Studies, 168*, 665-692.

Pears, D. (1982). How easy is akrasia? *Philosophia*, *11*(1-2), 33-50.

Pistole, J. S. (2012, August 16). TSA: Behavior detection is simply common sense. Retrieved from http://usatoday30.usatoday.com/news/opinion/story/2012-08-16/behavior-detection-TSA-Pistole/57102024/1

Posner, M. I., & Snyder, C. R. (2004). Attention and cognitive control. In D. A. Balota & E. J. Marsh (Eds.), *Cognitive psychology: Key readings* (pp. 205-223). New York: Psychology Press.

Posner, M. I., & Snyder, C. R. (1975). Attention and cognitive control. In *Information processing and cognition: The Loyola symposium* (pp. 55-85). New York: Halsted Press.

Pugmire, D. (1969). 'Strong' self-deception. *Inquiry, 12*(1-4), 339-346.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology, 46*(2), 237-248.

Raz, J. (2010). Responsibility and the negligence standard. *Oxford Journal of Legal Studies*, *30*(1), 1-18.

Rosen, G. (2003). Culpability and ignorance. *Proceedings of the Aristotelian Society, 103*(1), 61-84.

Rothschild, D. M. (2012, July 18). The Signal prediction: Obama maintains fragile hold on 303 electoral votes [Web log post]. Retrieved from http://www.predictwise.com/node/615

Rueda, M. R., Posner, M. I., & Rothbart, M. K. (2004). Attentional control and self-regulation. *Handbook of self-regulation: Research, theory, and applications*, *2*, 284-299.

Ruz, M., & Lupianez, J. (2002). A review of attentional capture: On its automaticity and sensitivity to endogenous control. *Psicologica, 23*, 283-309.

Sartre, J. (1943). *Being and nothingness*. London: Routledge.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs, 36*(2), 249-275.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *The Philosophical Review, 117*(2), 245-273.

Scott-Kakures, D. (1996). Self-deception and internal irrationality. *Philosophy and Phenomenological Research*, *56*(1), 31-56.

Scott-Kakures, D. (2002). At "permanent risk": Reasoning and self-knowledge in self-deception. *Philosophy and Phenomenological Research, 65*(3), 576-603.

Silver, N. (2012, July 27). July 27: Ohio polls show trouble for Romney [Web log post]. Retrieved from http://fivethirtyeight.blogs.nytimes.com/2012/07/27/july-27-ohio-polls-show-trouble-for-romney/

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059-1074.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*(1), 229-243.

Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., & Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biology, 5*(6), E138.

Smith, H. (1983). Culpable ignorance. *The Philosophical Review, 92*(4), 543-571.

Sosa, E. (1991). *Knowledge in perspective: Selected essays in epistemology*. Cambridge: Cambridge University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-665.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.

Sverdlik, S. (1993). Pure negligence. *American Philosophical Quarterly*, *30*(2), 137-149.

Teasdale, J. D., Segal, Z. V., Williams, J. M., Ridgeway, V. A., Soulsby, J. M., & Lau, M. A. (2000). Prevention of relapse/recurrence in major depression by mindfulness-based cognitive therapy. *Journal of Consulting and Clinical Psychology, 68*(4), 615-623.

Tierney, J. (2014, March 24). At airports, a misplaced faith in body language. Retrieved from http://www.nytimes.com/2014/03/25/science/in-airport-screening-body-language-is-faulted-as-behavior-sleuth.html

Tornello, D. N. (2014, April 8). NYT: TSA's behavior detection program useless, wasteful [Web log post]. Retrieved from http://tsanewsblog.com/13388/news/nyt-tsas-behavior-detection-program-useless-wasteful/

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. New York: Basic Books.

United States of America, Government Accountability Office (2013). *TSA should limit future funding for behavior detection activities* (pp. 1-93). GAO.

Vargas, M. (2004). Responsibility and the aims of theory: Strawson and revisionism. *Pacific Philosophical Quarterly, 85*(2), 218-241.

Vargas, M. (2006). On the Importance of history for responsible agency. *Philosophical Studies, 127*(3), 351-382.

Voorhees, J. (2012, November 6). Nate Silver says Obama has a better than 90 percent chance of winning the election [Web log post]. Retrieved from http://www.slate.com/blogs/the_slatest/2012/11/06/nate_silver_nyt_polling_expert_gives_obama_a_big_edge_on_election_day.html

Wallace, R. J. (1996). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Wang, S. (2012, November 6). Presidential prediction 2012 – final [Web log post]. Retrieved from http://election.princeton.edu/2012/11/06/presidential-prediction-2012-final/

Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*(8), 205-220.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M. (2005). Who is the controller of controlled processes? In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 19-36). New York: Oxford University Press.

Weiner, R. (2012, November 6). Mitt Romney hasn't written concession speech. Retrieved from http://www.washingtonpost.com/blogs/post-politics/wp/2012/11/06/mitt-romney-hasnt-written-concession-speech/

Written testimony of TSA Administrator John Pistole for a House committee on homeland security, subcommittee on transportation security hearing titled "TSA's SPOT program and initial lessons from the LAX shooting" (2013) (testimony of John S. Pistole).

Zimmerman, M. J. (1986). Negligence and moral responsibility. *Noûs, 20*(2), 199-218.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics, 107*(3), 410-426.