HOW MANY PERSONS AM I? AN EMPIRICAL INVESTIGATION OF PERSONHOOD AND THE UNCONSCIOUS

by

KAREN L. SHANTON

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Alvin Goldman

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

OCTOBER, 2011

ABSTRACT OF THE DISSERTATION

How Many Persons Am I? An Empirical Investigation of Personhood and the Unconscious

By KAREN L. SHANTON

Dissertation Director:

Alvin Goldman

Many philosophical positions depend on claims about the mind. Though it's tempting to think that the claims that matter – at least from a philosophical perspective – are claims about the *conscious* mind, emerging evidence suggests that the unconscious plays a surprisingly significant role in our mental lives. Given the centrality of claims about the mind to philosophical positions, and the centrality of the unconscious to the mind, it's important for philosophers to take account of discoveries about the unconscious. My dissertation is an attempt to do this. I use empirical findings about unconscious states and processes to investigate the nature of personhood and the relationship between human beings and persons.

Most contemporary empirical work on unconscious states and processes is conducted in two overlapping literatures: (1) the dual process literature and (2) the cognitive unconscious literature. According to the dual process literature, we have two

different ways – one conscious and the other unconscious – of performing many types of cognitive tasks. The central moral of the cognitive unconscious literature is that we underestimate the unconscious; unconscious processes are capable of much more sophisticated and flexible processing than we tend to think.

I draw on these overlapping literatures to challenge two common philosophical assumptions. In Chapters 1-2, I use findings in the dual process literature to challenge the assumption that there's a single, unified person in each human being. In Chapters 3-5, I use findings in the cognitive unconscious literature to challenge the assumption that consciousness is necessary for personhood. These two challenges combine to form a larger project. As I argue in the Coda, they raise the possibility that there are two distinct persons in each human being – one conscious and the other unconscious. This possibility, in turn, has implications for a range of philosophical issues, from diachronic personal identity to moral responsibility to animal and artificial consciousness.

ACKNOWLEDGMENTS

First and foremost, I owe thanks to Jonathan Haidt, my sister, Julia Lehman, and the snowy state of Minnesota. Had Julia not given me a copy of Haidt's *The Happiness Hypothesis* to read on a snowbound Minnesota day a few winters ago, I might never have found a dissertation topic. I undoubtedly wouldn't have found this one. So, thank you for helping start me on this path.

Many professors are hesitant to supervise dissertations on topics that aren't in their immediate research areas. Even more would have been skeptical of a dissertation about the unconscious. Fortunately, I had one of the most open-minded, generous, and supportive dissertation advisors in the business. Starting from my first semester at Rutgers, Alvin Goldman was an incredible mentor to me. His guidance, in matters of both style and substance, was truly invaluable.

I was also lucky to have a kind, constructive, and whip-smart dissertation committee. To Brian McLaughlin who kept me focused on the big picture, Frankie Egan who pushed me to ever greater precision and specificity, and Ned Block who inspired me to always keep learning more science, thank you. Whatever its current merits, my dissertation would have been a far inferior piece of work without your contributions. Thanks also to Barry Loewer and Holly Smith for your immensely helpful critiques and suggestions, and to Alan Leslie for welcoming me into your cognitive development lab.

The Rutgers graduate community was a lovely place to spend a few years. For their insightful comments, their friendship, or both, I'd like to thank Marcello Antosh, Tim Campbell, Heather Demarest, Kate Devitt, Tom Donaldson, Richard Dub, E.J. Green, Gabe Greenberg, Angela Harper, Michael Johnson, Lucy Jordan, Stephanie Leary,

Karen Lewis, Kelby Mason, Katy Meadows, Zak Miller, Lisa Miracchi, Alex Morgan, Jenny Nado, Ron Planer, Mary Salvaggio, James Simmons, Meghan Sullivan, and Jenn Wang. I would also be remiss not to express my appreciation to our wonderful graduate program administrators, Mercedes Diaz and Susan Viola.

At the top of a hill in the middle of Ohio, surrounded by acres of cornfields, sits an impossibly charming college campus. Kenyon College is the very definition of a liberal arts haven, and I was fortunate to spend four of the best years of my life there. My professors at Kenyon were some of the finest teachers our educational system has to offer. To Fred Baumann who challenged and encouraged me in equal measure, Andrew Pessin who guided me through my first large-scale philosophy project, and Joel Richeimer who showed me that you can sneak cognitive science into anything (even history of ancient philosophy), thank you. Thank you for all the time and energy you shared with me and the generations of lucky Kenyonites who came before and after me.

Philosophers often muse about how lucky we are to get paid to do what we do. Though perhaps cliché, this couldn't be more true. Thank you to the Andrew W. Mellon Foundation, the Jacob K. Javits Fellowship Program, the Rutgers Graduate School of Arts and Sciences, and the Rutgers Department of Philosophy for your generous financial support.

Last, but certainly not least, I would like to thank my parents, Ken and Audrey Shanton, my sister and brother-in-law, Julia and Vance Lehman, and my beautiful niece, Elise. Words can't adequately express how grateful I am for the encouragement you've offered me over the years. So, let me just say thank you and I love you. And your copies of this dissertation are in the mail. There'll be a quiz on it in the morning.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

Many philosophical positions depend on claims about the mind. Think, for example, about work on diachronic personal identity. Arguably the most popular approach to diachronic personal identity is the psychological continuity approach. According to psychological continuity theories, what makes *x* at t_1 the same person as *y* at t_2 is some sort of psychological continuity between *x* and *y*. What kind of psychological continuity does there have to be? That depends (at least in part) on what the mind is like. Suppose, for example, that Puccetti (1993) is right about the morals we should draw from work with split-brain patients (patients who have had the bundle of nerves connecting the right and left hemispheres of their brains severed). Split-brain patients seem to have two distinct streams of thought (Sperry et al., 1969; Sperry, 1974). For example, a split-brain patient might button his shirt with one hand while unbuttoning it with the other, or attempt to hit his wife with one hand while holding himself back with the other (Bogen, 1993; Geschwind, 1981). Puccetti takes this to show that there are two separate persons in split-brain patients. More generally, he thinks, it shows that there are two persons in *all* human brains. Suppose he's right about this. An implication would be that psychological continuity theorists should be looking for two chains of psychological continuity in each brain rather than just one.

It's tempting to think that the important claims about the mind – at least from a philosophical perspective – are claims about the *conscious* mind. After all, the two streams of thought in split-brain cases are often described as dual streams of *consciousness*. More generally, we seem to be able to account for our experience of, and interaction with, the world pretty much entirely in terms of conscious states and

processes. Conscious perceptions seem to give us fairly complete access to the external world, and we seem to be able to explain our thoughts and behaviors in terms of conscious beliefs, desires, etc. This might lead us to think that the really important parts of the mind – the ones we *really* need to consider when developing philosophical positions – are the conscious parts. This, in turn, might lead us to focus our philosophical investigations into the mind on conscious states and processes. When specifying the continuities that are required for diachronic personal identity, for example, we might think that we can limit our focus to *conscious* continuities.

Limiting ourselves in this way would be a mistake. Think about your daily life. Think, for example, about the last time you attempted a *New York Times* crossword. If your crossword experiences are anything like mine, the following should be familiar. One of the clues – say, 52-across – stumps you when you first read it. After trying to think it through a time or two, you move on to other clues. As you're working on other parts of the puzzle, the answer to 52-across suddenly flashes into your mind. Alternatively, suppose you're driving in an area you've only visited a time or two before. You suddenly realize that you're lost (and have left your trusty GPS at home). Though you aren't completely certain of it, you have a hunch that you should take a right at the next street. Cases like these – flashes of insight and hunches – hint at unconscious influences on our thoughts and behavior.

Recent empirical work confirms this undercurrent of unconscious influence. For one thing, studies suggest that we don't have the conscious processing capacity to produce all (or even most!) of our thoughts and behaviors. Indeed, conscious processes seem to play a causal role in only about 5% of behavior production (Baumeister et al.,

1997; Baumeister et al., 1998). Unconscious thought and behavior production has also been directly experimentally demonstrated. Perhaps the most famous of these demonstrations is Nisbett & Wilson's (1977) work on decision-making. Participants in Nisbett & Wilson's study were presented with four identical products, told to choose one of them, then asked why they made the choice they did. Analyses of the overall data revealed that participants were susceptible to a position effect; they were more likely to choose products on the right than the left. When asked why they chose the product they did, however, none of the participants spontaneously cited the product's position as a reason for their choice. Indeed, even when *explicitly* asked whether position played a role in their decisions, they firmly denied that it did. This finding not only confirms that thoughts and behaviors can be produced unconsciously, but also shows that such unconscious influences aren't limited to cases (like the crossword puzzle and driving cases) in which they make themselves known to us; unconscious influences often operate *completely* outside our conscious awareness.

Given the centrality of claims about the mind to many philosophical positions and the centrality of the unconscious to the mind, it's important for philosophers to take account of discoveries about the unconscious. My dissertation is an attempt to do this. In recent years, there's been a veritable avalanche of new – often surprising – discoveries about the unconscious and its role in cognition. My aim is to identify some of the implications of these discoveries for philosophy. More specifically, I try to show that the discoveries challenge some common philosophical assumptions about personhood. I start, in this Introduction, by explaining exactly what I mean by 'conscious' and 'unconscious,' and giving a more detailed overview of my projects.

I.1 Conceptualizing Consciousness

There's a special challenge that confronts researchers who work on interdisciplinary issues. Call this the 'false friends problem.' If you've ever studied a Romance language, you've almost certainly encountered some pairs of false friends. False friends are words that look or sound similar, but have different meanings. One example of such a pair is *embarrassed* and *embarazada*. Though the Spanish word *embarazada* looks and sounds quite a bit like the English word *embarrassed*, it doesn't actually *mean* 'embarrassed'; rather, it means 'pregnant.' Similar phenomena can occur across disciplinary lines; researchers in different disciplines sometimes use words that look or sound the same, but have different meanings.

When we try to apply empirical findings to philosophical issues, then, we have to make sure we're aware of – and account for – the false friends problem; we have to make sure that the philosophical and scientific usages of the terms we're studying line up with each other. There are two ways to go about this. One is to start with the philosophical usage, and look for a scientific term that matches it. The other is to start with scientific usage, and look for a philosophical match. Which of these methods we should use depends on the nature of our projects. My project is to apply empirical findings to philosophical issues. Therefore, the better bet for me is the second strategy; I should start with scientific usage then match it to philosophical terms.

So, let's begin by asking what brain scientists mean when they talk about the conscious and the unconscious. There have been times – most notably the Freudian era – when 'unconscious' had a loaded meaning. For Freud, the unconscious was a product of active repression. Painful, socially unacceptable, and traumatic thoughts were pushed

down into an unconscious mind. From there, they battled for expression with forces in the conscious mind. This picture has not, however, retained its popularity. Due largely to concerns about the falsifiability of Freud's complex system, brain scientists have generally moved away from this model of the mind (Neuroskeptic, 2009). More recent references to the unconscious – like the ones I discuss in this work – tend to offer a more minimal account of it. According to this minimal account, the unconscious is just the part of the mind that is not accessible to introspection.

What exactly does it mean to be inaccessible to introspection? More fundamentally, what's 'introspection'? James thinks the answer here is obvious: "The word introspection need hardly be defined – it means, of course, the looking into our own minds and reporting what we there discover" (1890/1981, p. 85). Gertler (2009) amplifies this definition, explaining that the 'looking' James has in mind is not visual perception; rather, it's a kind of self-directed attention. There's some debate about exactly what form this self-directed attention takes. Some think of it as a quasi-perceptual process while others believe there's a more direct connection between introspection and its objects (Gertler, 2009). For present purposes, we don't need to settle this debate. What matters for now is just that there is a self-directed attention process by which we can access our mental states. This self-directed attention process is introspection, and it marks the difference between consciousness and unconsciousness. The conscious part of the mind is the part that can be accessed using this process while the unconscious is the part that can't.

How can we tell that accessibility to introspection is what contemporary brain scientists mean by 'consciousness'? One way is by looking at what they say when they

talk explicitly about consciousness. Take, for example, the discussion of conscious and unconscious processes in Uleman's (2005) historical overview of the dual process theory literature.[1] Uleman notes that dual process theorists draw a distinction between conscious processes and unconscious processes, and he cashes out this distinction in terms of (in)accessibility to introspection.

Another way to tell what brain scientists have in mind is by looking at the measures they use to test for consciousness. One common measure is experiential report. When studying unconscious perception, for example, psychologists often ask participants to report whether they see a stimulus. This seems to call for participants to introspect their visual perceptions. Another common set of measures is behavioral. To test whether young children were conscious of a particular piece of information, Ruffman et al. (2001) tracked their betting behavior. The logic behind using this measure was that the sizes of bets co-vary with confidence about them, and confidence about bets co-varies with introspectively accessible information about them. If you have introspective access to information that confirms a prediction, you'll be more confident that the predicted outcome will come to pass, and the more confident you are that a prediction will come to pass, the more you'll be willing to bet on it. Like experiential reports, this kind of measure seems designed to assess introspectibility.

Now, introspection is not an entirely uncontroversial topic. Before trying to match this scientific understanding of consciousness to philosophical usage, let's pause to address some of the controversies surrounding it. Carruthers (2009, 2010) – echoing earlier arguments by Ryle (1949) – contends that there are some types of mental states we can't introspect. Specifically, he thinks we don't have introspective access to any of our

[1] For a detailed discussion of this literature, see Chapters 1-2.

propositional attitudes. Instead, he claims, we access our *own* propositional attitudes the same way we access *others'* propositional attitudes: by interpreting evidence. We tend to be better at identifying our own propositional attitudes than others' attitudes because we have more evidence to draw on in the first- than the third-person case. However, the *methods* we use in both cases are the same. One of his arguments for this conclusion is that there is an evolutionary story to be told about why first- and third-person mental state attribution methods would overlap and no similar story about the development of introspection. He also appeals to parsimony and evidence of paired deficits in first- and third-person mindreading in individuals with autism. Finally, he cites the absence of compelling evidence for the use of introspection in accessing propositional attitudes.

As Carruthers (2009, 2010) himself would likely concede, the first two of these arguments are far from conclusive. A compelling evolutionary story can be told about the development of almost anything – indeed, Schulz (2010) tells one about the evolution of introspective mindreading – and parsimony only comes into play in theory choice when the theories under consideration explain the data equally well. The second two arguments don't fare much better. First, the paired deficit evidence Carruthers cites is evidence for paired deficits in third-person mindreading and first-person *past* mindreading. As Goldman (2006) emphasizes, however, no one claims that we use introspection for first-person *past* mindreading. In fact, Goldman and Shanton (Goldman & Shanton, forthcoming; Shanton & Goldman, 2010) have developed a non-introspective – specifically, a *simulationist* – account of first-person past mindreading. Therefore, evidence regarding first-person past mindreading is irrelevant to the introspection debate. Second, *pace* Carruthers, there *is* compelling evidence for the use of introspection in

accessing propositional attitudes. In an unpublished manuscript, Goldman cites cases of first-person propositional attitude attribution that can't readily be explained by Carruthers' interpretive account. Suppose, for example, that you are trying to determine whether you intend to take a vacation in July. There's no obvious interpretive route to this determination. What evidence could we use to arrive at a self-attribution here? A more plausible story is that we retrieve the intention from memory then introspect it.

A related (though less radical) controversy about introspection centers on the methods used to test for it. The objection here isn't that individuals don't use introspection to access their mental states. Rather, it's that the methods by which researchers *test* for the use of introspection are flawed. The primary concern here is about the use of experiential report. In order to report an introspected experience successfully, you have to both introspect the experience and verbalize your introspection. This raises the possibility that failures to report experiences might be due to *verbalization* failures rather than *introspective* ones. Participants might fail to report their experiences not because they can't introspect them, but because they can't verbalize them.

There are a couple of ways to respond to this kind of challenge. One is to emphasize behavioral measures of introspectibility instead of (or in addition to) experiential report. Behavioral measures aren't always well-equipped to establish the *presence* of consciousness; unconscious states and processes can produce behaviors, so the presence of a behavior isn't often evidence for the presence of consciousness. However, they are consistently good at establishing the *absence* of consciousness (or the presence of unconsciousness). If a participant fails to perform certain behaviors, we can infer that he lacks introspective access to the states or processes that would prompt those

behaviors. For an example of how this works, think about Ruffman et al.'s (2001) use of the betting paradigm. In betting cases, introspective access to disconfirming evidence would lead to tempered betting behavior. So, if betting behavior is *not* tempered in response to disconfirming evidence, we can infer that the bettor doesn't have introspective access to the disconfirming evidence.

Another way to respond to the challenge is by appealing to a strategy initially proposed by Merikle & Daneman (1999). Merikle & Daneman noticed that attempts to establish the existence of unconscious perception stalled over disagreements about what count as genuine criteria of introspectibility; no matter what criterion was proposed, objections could be raised to it. They suggested that, instead of trying to arrive at a consensus about criteria of introspectibility, we should do something different. We should divide processes into two groups on the basis of experiential report then check whether there are qualitative differences between the processes in the two groups. If there are, we can infer that there is a genuine, fundamental difference between the two sets of processes. When Merikle & Daneman tested this strategy on reportable and unreportable perceptions, they found that there *were* qualitative differences between the two sets.[2] Therefore, the reportable / unreportable distinction does seem to be tapping into an important difference between processes.

Now, neither of these responses is bulletproof. Challenges like the challenge raised to experiential report could be raised to behavioral measures. It could also be objected that evidence that there is a genuine, fundamental difference between reportable and unreportable processes is not evidence that the difference is a difference in

[2] As we'll see in Chapter 1, this finding is not unique to perception. There are similar qualitative differences between the reportable and unreportable versions of many other types of processes.

introspectibility. The inability to provide conclusive proof of (a lack of) introspectibility is, however, an unavoidable consequence of the nature of the phenomena we're trying to study and the current state of our technology. Unless (or until) we can get inside participants' heads, we just can't know for sure what they can and can't introspect. While we wait on the monumental advances in technology this would require, we'll have to make do with the methods we do have available. In the absence of positive reasons to think that these methods don't measure introspectibility – rather than speculations that they *might* not – let's proceed on the assumption that they do.

In the preceding, I've talked a lot about 'the unconscious.' This locution carries a connotation of unity; it implies that 'the unconscious' is a unified system. This isn't necessarily a false connotation. As we'll see in Chapter 5, there's reason to believe that the unconscious *is* a unified system. When I talk about 'the unconscious,' however, I don't mean for it to carry this connotation. This phrase should be read more neutrally, as referring just to the set of states and processes that are inaccessible to introspection.

What are these states and processes? It should be clear what I mean when I say that a mental *state* is inaccessible to introspection. An introspectively inaccessible mental state is simply a mental state that can't be introspected. Defining introspectively inaccessible *processes*, on the other hand, isn't quite as straightforward. Mental processes have more moving parts than mental states. They're made up of:

(1) Inputs

(2) Outputs

(3) Mechanisms that translate inputs to outputs

Any or all of these parts could be inaccessible to introspection. A process might be kicked off by an introspectively inaccessible input – like one of the unconsciously perceived stimuli Merikle & Daneman (1999) investigate – or give rise to an introspectively inaccessible output. The mechanisms that connect the inputs to a process to its outputs might be inaccessible to introspection, as in the crossword puzzle case. Which of the parts of a process have to be introspectively inaccessible for the process as a whole to count as unconscious?

We can characterize unconscious process types either broadly or narrowly. Types of processes tend to be individuated in terms of the mechanisms they employ. Compare, for example, using a multiplication table and actually working out the multiplication. These two types of processes can take the same inputs (e.g. 12 x 12) and produce the same outputs (e.g. 144). What makes them different types of processes is that they use different *mechanisms* to translate the inputs into outputs. Similarly, one way to differentiate conscious from unconscious processes is solely in terms of mechanisms: unconscious processes are processes whose mechanisms are completely inaccessible to introspection while conscious processes are processes whose mechanisms are not completely introspectively inaccessible.[3] This solely-mechanism-based characterization of unconscious processes is what I describe as the broad characterization of unconscious processes. The narrow characterization is a bit more demanding. According to this characterization, a process is unconscious only if both its mechanisms *and* its inputs are unconscious. The broad characterization of unconscious processes is my default

[3] To understand the difference here, compare a case in which you work through a crossword puzzle clue to one in which the answer comes to you in a flash of insight. In the first case, steps in the process that translates inputs to outputs are accessible to introspection while, in the second, they aren't.

characterization of unconscious processes; unless I flag that I intend the narrow characterization (as I will do in Chapter 3), I am assuming the broad characterization.

How does the scientific understanding of consciousness we've been considering line up with philosophical usage of the term? Introspective accounts of consciousness are popular in philosophy. Introspection tends to be what higher-order perception theorists are talking about when they talk about consciousness and, arguably, it's also the core of our commonsense understanding of the term. When we talk about conscious beliefs, for example, we typically mean *beliefs to which we have introspective access*. Similarly, unconscious desires are desires we didn't (in a certain sense) realize we had. They are desires we can't directly introspect.

Introspective consciousness is also largely coextensive with Block's (1995) phenomenal consciousness. Block has famously drawn a distinction between phenomenal consciousness and access consciousness. A mental event is phenomenally conscious provided there is something it is *like* to be in it (Nagel, 1974). Suppose you have the bad fortune to be trapped in a real-life version of a horror movie scenario. The fear you feel when you realize the killer's call *is coming from inside the house* is a phenomenally conscious state. A mental event is access conscious provided that it is poised for use in rational thought and action. Suppose you use your belief that the killer is inside the house to form and execute an escape plan. In this case, your belief is access conscious.

There are a couple of different ways to think about phenomenal consciousness. Some philosophers – call them 'phenomenological liberals' – think all introspectible mental states, including introspectible propositional attitudes, have phenomenology (see, for example, Pitt, 2004). These philosophers argue that we couldn't introspect our

propositional attitudes unless there was something it was like to be in them. For example, you couldn't introspect your belief that the killer is inside the house unless that belief *felt* like something to you. Other philosophers – call them 'phenomenological conservatives' – think propositional attitudes don't have phenomenology (see, for example, Lormand, 1996). These philosophers argue that any 'feel' propositional attitudes seem to have is actually the feel of an imagistic representation that accompanies the attitude. Though your belief that there's a killer in the house is accompanied by a feel, for example, it isn't the feel of the belief itself. Rather, it's the feel of your visualization of the killer lying in wait at the bottom of the stairs or the words, 'There's a killer in the house.' Phenomenal consciousness in the first – liberal – sense overlaps with the introspective scientific understanding of consciousness.

I.2 A Set of Projects

The human mind is, in many ways, very different than we tend to think it is. A pair of overlapping scientific literatures – the dual process literature and the cognitive unconscious literature – suggests that unconscious states and processes are more pervasive in, and central to, human cognition than we tend to assume. According to the dual process literature, we have two different ways – one conscious and the other unconscious – of performing many types of cognitive tasks. Though we're only introspectively aware of the operations of the conscious process, either type of process can be responsible for our thoughts and behaviors. For example, it's well-documented that participants respond differently to structurally similar versions of ethical dilemmas like the trolley problem. The reason for this, dual process theorists argue, is that participants' responses to one version of the dilemma are produced by one (conscious)

process while their responses to the other are produced by a different (unconscious) process (Greene et al., 2001). Each of the processes in the dual process pair produces some of participants' moral responses.[4]

The central moral of the cognitive unconscious literature is that unconscious processes are capable of much more than we tend to give them credit for. Unconscious processes are typically thought to be pretty limited. Like physical reflexes, they're characterized as automatic and ballistic; they occur automatically in response to a certain range of stimuli and, once started, can't easily be stopped. Like assembly line workers, they're attributed only a limited range of responsibilities (and capabilities). They are trained for a certain, circumscribed set of tasks and, though they can get very good at these tasks, they can't go beyond them; they can't go beyond their training. Flexibly responding to changes in the environment or combining two pieces of information to solve a novel problem is, it's often thought, beyond the unconscious' abilities. For these kinds of tasks, we need consciousness. The cognitive unconscious literature upends this traditional thinking about the unconscious. Studies in this literature show that unconscious processes are much more sophisticated and flexible than we tend to think. Indeed, the list of tasks that can be performed unconsciously has got so long that it's prompted a leading cognitive unconscious researcher – Dijksterhuis (2009) – to muse that the real question isn't what the unconscious *can* do, but what it *can't* do.

These two literatures suggest a significant reimagining of both the nature of the mind and the relationship between the conscious and the unconscious. Given the centrality of claims about the mind to many philosophical positions, it seems improbable that such a substantial change could have no philosophical reverberations. We would

[4] For a more detailed discussion of Greene et al.'s (2001) work, see Chapter 1.

expect major revisions like these to ramify through many different areas of philosophy. In this dissertation, I trace some of these ramifications. More specifically, I trace the ramifications of the revisions for *personhood*.

I start, in Chapters 1-2, by drawing out the implications of dual process work for the assumption that there's a one-to-one relationship between human beings and persons. It's commonly assumed that – at least as far as normal adult human beings are concerned – there's one person per human being (call this the 'one man-one vote assumption'). The one man-one vote assumption is central, for example, to the case for the animalist approach to personal identity. According to animalism, you (the person) are numerically identical to a human animal. In motivating this kind of approach, Olson appeals to the (apparent) fact that "In every actual case, the number of people we think there are is just the number of human animals. Every actual case in which we take someone to survive or perish is a case where a human animal survives or perishes" (2003, p. 332). This is a bit hyperbolic. As we'll see shortly, there are *some* actual cases in which there seems to be something other than a one-to-one relationship between persons and human animals. Nonetheless, I think it accurately captures the general, commonsense take on the relationship between persons and human beings: exceptional cases aside, each human being contains a single person.

The one man-one vote assumption is also implicit in the way we individuate moral units. Persons seem to be basic units of moral consideration; persons are the kinds of things that can be morally responsible, have self-interests, etc. Now, when we make moral calculations, we tend to treat human beings as wholes, rather than parts or groups of human beings, as our units of moral consideration. This suggests that we typically take

human beings as wholes (vs. parts or groups of human beings) to be good stand-ins for persons. The fact that we use calculations about human beings as shorthand for calculations about persons suggests that we take there to be substantial overlap between the class of human beings and the class of persons.

Further evidence for a general commitment to the one man-one vote assumption is the way we respond to abnormal cases. Take, for example, the case of patients with multiple personality disorder (or, as it's now more commonly known, dissociative identity disorder). It's sometimes argued that the multiple personalities in dissociative identity disorder patients are distinct persons (see, for example, Wilkes, 1988). That this is taken to require argument suggests that such multiplicities of persons are not the norm. In the ordinary case, it suggests, there aren't multiple persons lurking in any given human being. Only when there is something abnormal or unusual about a human being does he contain more than a single, unified person.

In Chapters 1-2, I argue that this one man-one vote assumption – widely accepted though it is – is challenged by the dual process findings. If we take the dual process work seriously, we end up with a picture of the human mind according to which there are multiple different forces, with multiple different personalities, competing for control over the mind (and body). If this picture is right, I contend, normal human beings don't contain a single, unified person. These kinds of deep psychological schisms aren't compatible with thinking of the divided mind as a single person. Instead, we should model individual normal human beings on groups of human beings. Just as we traditionally conceive of different personalities in *different* bodies as not-the-same-person, we shouldn't lump different personalities in the *same* body together.

In Chapters 3-5, I turn to questions about the function of consciousness. Consciousness is often thought to be *special*; it seems to enable us to do things we couldn't do without it. This (alleged) specialness is often taken to have metaphysical consequences. For example, Nelkin talks about "the cognitive consciousness that makes us persons" (1993, p. 234). The general idea here is that there are certain capacities (e.g. rationality) that are necessary for personhood, and these capacities are unique to conscious beings. To be able to perform these functions, you have to be conscious and, to be a person, you have to be able to perform the functions. Ultimately, then, the conscious-unconscious boundary is a bright line between beings that are candidates for personhood (conscious beings) and beings that aren't (unconscious beings).

As noted above, the general moral of the cognitive unconscious literature is that we shouldn't underestimate unconscious processes. This moral holds as firmly in the current context as it does in general. In Chapter 3, I lay out some candidates for functions of consciousness that are required for personhood (to anticipate a later bit of labeling, let's refer to these functions as 'metaphysical functions of consciousness'). For each candidate, I then show either that it isn't unique to consciousness – consciousness is, in fact, not required for performance of the function – or that it isn't actually required for personhood. This raises doubts about the specialness of consciousness – or, at least, the *personhood-relevant* specialness of consciousness. It's indisputable that conscious beings differ from unconscious beings. Arguably, they differ from unconscious beings in ways that make them special. If the argument in Chapters 3-5 goes through, however, these differences don't make them *metaphysically* special. More specifically, they don't make them candidates for personhood in a way that unconscious beings are not. The conscious-

unconscious boundary *isn't* a bright line between (conscious) candidates for personhood and (unconscious) non-candidates.

The arguments in Chapters 1-2 and Chapters 3-5 combine to form a larger project. In Chapters 1-2, I draw on dual process work to show that the normal human being isn't a single, unified person. Though it's tempting to equate this conclusion with the conclusion that there's more than one person in the normal human being, more work has to be done to establish this further conclusion. To show that there are actually multiple persons in the normal human being, we have to show not only that he is psychologically divided in ways that prevent him from qualifying as a single, unified person, but also that more than one of the divided parts counts as a person in its own right. Chapters 3-5 are an attempt to make this second point. Drawing on the cognitive unconscious literature, I argue that the unconscious is capable of the kinds of functions, and has the kinds of characteristics, we expect of persons. This suggests that the unconscious force that competes with the conscious force for control over the normal human being's thoughts and behaviors is not merely a random spoiler. Instead, it might be a separate, second person in the same human body.

CHAPTER 1: AN INTRODUCTION TO DUAL PROCESS THEORIES

1.1 The Dual Process Hypothesis

Suppose you are asked to play a game. The game is a computer simulation of a sugar factory, and your job in the game is to control the factory's output. To help you perform this job, you are given either verbal instructions about how the factory works or practice controlling the factory. Now suppose you are tested on your knowledge about the factory's operations. One test is verbal: you are asked explicit questions about how the factory works, and expected to answer verbally. The other is practical: you are asked to demonstrate your ability to control the factory by producing a particular sugar output. How do you think you'll perform on these tests? If you're like most people, your performance will depend on the training you received. If you were given verbal instructions about how the factory works, you will perform better on the verbal test than the practical one. If you were given practice controlling the factory, on the other hand, you'll do better on the practical test. If you're like most people, you'll also have different degrees of consciousness of your learning processes in each of the cases. You'll be consciously aware of (at least part of) your learning process in the verbal instruction case, but not in the practice case.

What accounts for these differences in test performance and conscious accessibility? According to Berry & Broadbent (1984), the differences reflect the use of two different types of learning processes:

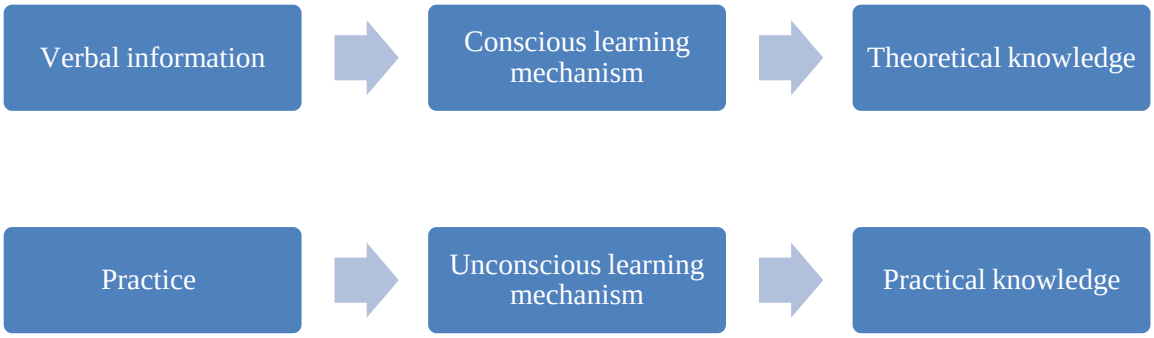

Figure 1

The learning mechanism in the first type of process takes verbal information about the simulated sugar factory as an input, and generates theoretical knowledge about the system as an output. The learning mechanism in the second type of process takes practice controlling the system as an input, and generates the ability to control the factory's sugar production (practical knowledge) as an output.

Now suppose you are presented with the following scenario. A girl decides to go outside to play. Before leaving the house, she puts her favorite toy in a box. While she is out playing, her brother comes into the room and moves the toy to a basket. When the girl comes back inside, she wants to find her toy. Where will she look for it? Where does she believe it is? Because she didn't see the toy being moved, the girl should expect it still to be where she left it (in the box). This is the belief normal adults attribute to the girl. Starting at around the age of four, it's also what young children say when asked. Before the age of four, however, children tend to say that the girl will look for her toy in the basket.

For many years, young children's failure verbally to pass this task (the false belief task) was taken to show that they aren't capable of mindreading, or mental state attribution (see, for example, Wimmer & Perner, 1983). However, more recent work adds

a wrinkle to the story. Onishi & Baillargeon (2005) found that children as young as fifteen months of age look longer at versions of the above scenario in which an agent looks for the object in the new location than versions in which she looks in the original location. Looking time is commonly accepted by developmental psychologists as a measure of surprise; the longer an infant looks at a scenario, the more surprising he finds it. Therefore, the fact that infants looked longer at the scenario in which the agent looks for the object in the new location suggests that they found this behavior surprising; they *expected* her to look for the object where she had left it (in the original location). To put this another way, the infants believed that the agent believed the object was where she left it. They correctly attributed to her the false belief that the object was in the original location.

How can we explain the fact that young children seem to attribute the correct belief to the agent in these scenarios when their responses are measured nonverbally (using looking times), but the incorrect belief when their responses are measured verbally? An earlier study, using the verbal version of the false belief task, suggests an answer. Prior to Onishi & Baillargeon's (2005) development of the nonverbal version of the false belief task, researchers using the verbal version noticed something strange: some preschoolers (3-5-year-olds) who gave the wrong *verbal* response to the task passed it 'with their eyes.' If the correct response was that the girl would look for her toy in the box, for example, these children *said* she would look in the basket but *looked at* the box (Clements & Perner, 1994). There are two possible explanations of these eye gaze data. First, the data might reflect low-confidence *conscious* awareness of the correct answer. The children might be consciously aware of the right answer, but not completely

convinced of it. Second, they might reflect *unconscious* awareness of the correct answer. Participants might be unconsciously, but not consciously, aware of the girl's false belief.

Using a betting paradigm, Ruffman et al. (2001) tested these two explanations. Betting behavior is sensitive to uncertainty; people tend to be unwilling to bet on responses of which they are unsure. So, if the eye gaze data reflected low-confidence conscious awareness of the correct answer, we would expect children to bet only moderately on their verbal responses. Their confidence in the verbal response would be tempered by their conscious awareness of the alternative possibility, and this would drive down their bets. If the eye gaze data reflected unconscious awareness of the correct answer, on the other hand, betting behavior should be unaffected by it; the children should be willing to bet a fair amount on their answers. Ruffman et al.'s findings favor the second of these explanations. Participants in their study bet strongly on their verbal responses.[5]

These findings suggest the following possibility: the children in Clements & Perner's (1994) and Ruffman et al.'s (2001) studies were employing two different types of mindreading processes. Here's roughly how these processes might look:

[5] Betting might seem too sophisticated a behavior for preschoolers to engage in. However, this turns out not to be the case. Before they started testing, Ruffman et al. (2001) had their participants complete a training exercise. In this exercise, they taught them that counters bet on correct answers would be doubled while counters bet on incorrect answers would be lost. They motivated them to try to acquire as many counters as possible by showing them a sheet that (ostensibly) listed the best performances by other participants. Analysis of participants' performance on this training exercise suggests that participants had little difficulty either picking up the betting paradigm or being motivated to perform well on it.

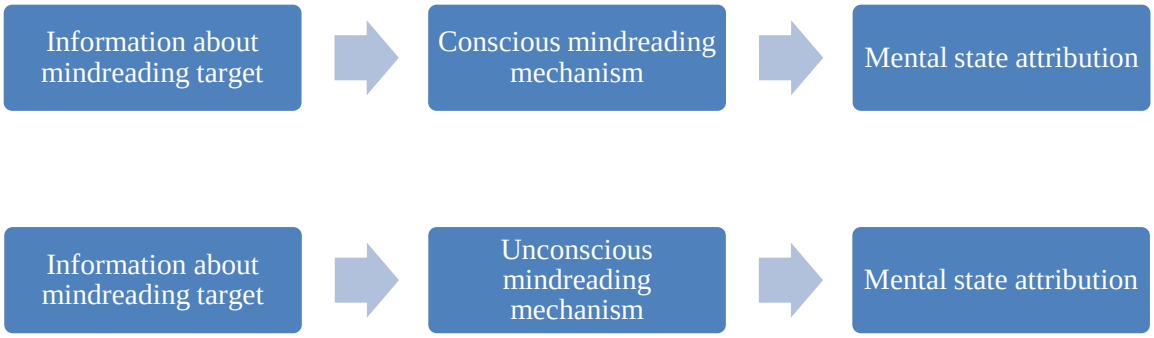

Figure 2

In the false belief task case, both mindreading mechanisms take information about the girl's movements, what she was able to see, what happened to the toy while she was outside, etc. as inputs. One of them then generates attribution of the belief (expressed in the child's verbal response) that the toy is in one location while the other outputs attribution of the belief (expressed in the looking response) that it is in a different location.

> Finally, suppose you're presented with the following case:
>
> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love? (Haidt, 2001, p. 2).

If you're like most people, you think it wasn't ok for Julie and Mark to have this sexual experience. When asked why, you might express concerns about the dangers of inbreeding or the emotional harm Julie or Mark will suffer. However, these concerns are obviously unfounded. Julie and Mark used two forms of birth control, so pregnancy was

unlikely.[6] The passage also makes clear that neither Julie nor Mark suffered – either emotionally or physically – from the sexual experience. If anything, they benefited from it. Given that these concerns are unfounded, why do you cite them as reasons? If it's clear that neither Julie nor Mark was harmed by the experience, why would you claim that your judgment about the case was motivated by your concern for their well-being?

Haidt (2001) offers the following explanation: there are actually two separate justification-generating processes at work here. We tend to think that reasoning about cases like Haidt's incest case looks something like this:

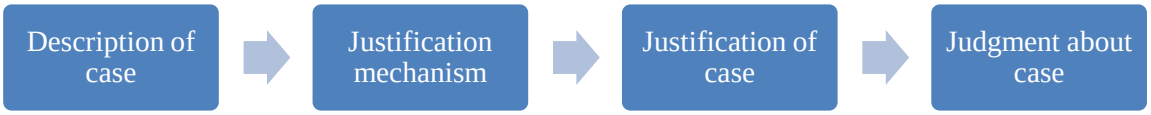

Figure 3

According to this model, when presented with a case, we generate reasons for and against possible responses to it, and weigh these reasons against each other to come up with a justification for one of the responses. This justification then forms the basis for our judgment about the case. Though intuitive, this model doesn't fit well with responses to cases like Haidt's incest case.[7] Haidt, therefore, proposes an alternative model of our reasoning in such cases. This model looks something like this:

[6] I would wager that the case could also be rewritten explicitly to rule out pregnancy (e.g. by making Julie, Mark, or both incapable of reproducing) without much change to your response to it.

[7] If this were the correct model of our moral justification-generating process, we shouldn't have any trouble providing plausible justifications for our judgments. Therefore, the fact that participants weren't able to provide plausible justifications suggests that there isn't a clear, direct, linear path from awareness of the case through conscious justifications to a judgment (as in this model).

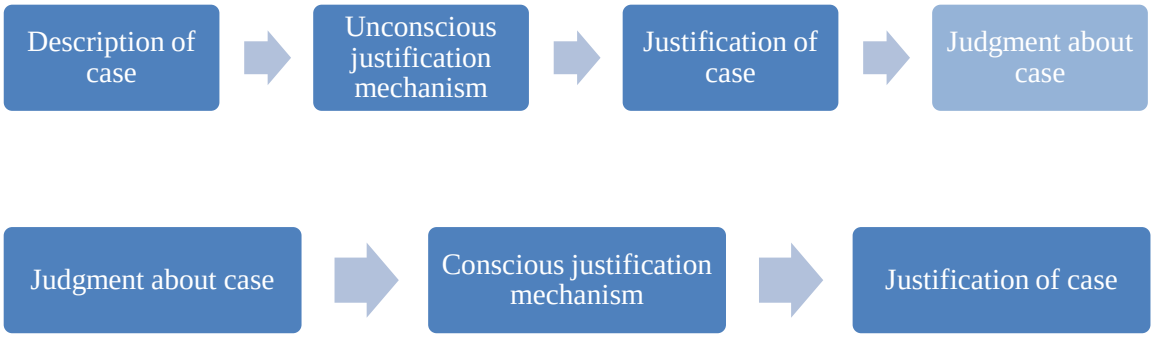

Figure 4

According to this model, your judgment about cases like the incest case is driven by an emotional – specifically, a *disgust* – response to the case. You don't have conscious access to this response or the process by which it generates your judgment, so you don't consciously *know* why you render the judgment you do. Therefore, when pressed to explain your judgment, you *confabulate* an answer. You fabricate a *post hoc* justification.

On this model, your response to the incest case involves two separate moral justification processes: an unconscious justification process and a conscious justification process. The unconscious justification process is the one that ultimately leads to your judgment about the case while the conscious justification process just provides a *post hoc* rationalization of this judgment. Interestingly, use of two different moral justification processes doesn't seem to be unique to the incest case; Haidt et al. (1993) also found similar patterns of responses to other cases (e.g. the case of a family eating the family dog after it was run over by a car).

The above three theories – of learning, mindreading, and moral justification – are examples of what psychologists refer to as dual process theories. Dual process theories are theories that instantiate the following hypothesis:

Dual process hypothesis: For a given type of cognitive task, *t*:

(1) There are two types of processes that can perform *t*.

(2) Each of these types of processes performs *t* in a different way.

Each part of this hypothesis could use further elaboration. According to the first part, the processes in dual process pairs both perform the same type of cognitive task. But what does it mean for two tasks to be of the same type? What characteristics do they have to share to be grouped together? According to the second part, the processes perform the same type of task in different ways. But there's more than one way in which task performance methods could differ. In which of these ways do *dual processes'* task performance methods differ? In §1.2 and §1.3, respectively, I address these two sets of questions.

1.2 Sameness of Types of Tasks

What makes two tasks tasks of the same type? A prior question is: what does it mean for something to be a 'task' at all? Tasks are things to be accomplished. If you ask your spouse to perform a cleaning task or your assistant to perform a filing task or your mechanic to perform a repair task, you are asking them to *accomplish* something (cleaning part of the house, filing a document, or fixing your car). This suggests that we should think of tasks *teleologically*; tasks are defined in terms of the ends toward which they are directed. If *tasks* are defined teleologically, it makes sense also to define *types of tasks* teleologically. Here, then, is a basic criterion of sameness of types of tasks:

Teleological criterion: Two tasks are of the same type provided that they are directed at the same type of end.

Applying this criterion is more difficult than it looks. Why? Because there's rarely only one way in which ends can be similar to or different from each other. The ends of cognitive tasks, in particular, can be similar or dissimilar in any or all of three ways:

(1) Mental state type

(2) Subject matter

(3) Domain

To illustrate these dimensions of similarity and dissimilarity, let's consider some examples. Suppose we have three pairs of tasks. One of the tasks in the first pair is aimed at producing *beliefs* while the other is aimed at producing *desires*. One of the tasks in the second pair is aimed at generating beliefs about a proposition, *p*, while the other is aimed at generating beliefs about a different proposition, *q*. Finally, one of the tasks in the third pair is directed at producing a *moral* belief while the other is directed at producing a *practical* belief. The tasks in these pairs differ along the first, second, and third of our dimensions. The pairs of tasks are aimed at generating mental states of different types, with different subject matters, and in different domains, respectively.[8]

Which of these dimensions of potential similarity and dissimilarity are relevant to determining whether two tasks are of the same type? The answer to this question depends on the purpose for which we are trying to identify tasks of the same type. At the moment, we're trying to explain why the three theories detailed in §1.1 are all grouped under the same heading. Why are these three theories – along with a host of others – all given the

[8] We can think of domains as categories of mental state types. For example, consider the belief that I should eat the family dog. This can be a moral belief (i.e. the morally appropriate thing to do is eat the family dog) or a practical belief (i.e. it's in my best practical interest to eat the family dog). These beliefs are of the same mental state type (belief) and have the same subject matter (I should eat the family dog) but are in different domains (moral vs. practical).

label ‘dual process theory’? In what way(s) do the outputs of a pair of processes have to be similar for the processes to qualify as *dual* processes?

They don’t seem to have to be in the same domain. Consider Berry & Broadbent’s (1984) learning processes, for example. As the diagrams in §1.1 show, the first of these processes outputs *theoretical* knowledge while the second outputs *practical* knowledge. Similarity of subject matter doesn’t seem to be necessary for dual processes either. Think, for example, about the justification-generation processes posited by Haidt (2001). Those processes are taken to act on stimuli with a range of different subject matters – from incest to family dog consumption – and, thus, to generate outputs with a range of different subject matters. This subject matter generalizability is not unique to Haidt’s processes. Indeed, *most* dual process theorists take their processes to generalize beyond the specific subject matters they’ve tested. For example, though Berry & Broadbent focused on the operations of a computer-simulated sugar factory in their study, they clearly intend their conclusions to generalize beyond this specific subject matter. The processes they posit are supposed to be *general* learning processes, not *computer-simulated sugar factory operations-specific* learning processes.

So, what *does* matter for determining whether a theory satisfies the first part of the dual process hypothesis? The answer, it seems, is similarity of mental state type. The common denominator in all the sample dual process theories in §1.1 is mental state type similarity; the processes in each of those three pairs both generate outputs of the same mental state type. If what we want to know is whether a theory satisfies the first part of

the dual process hypothesis, then, we need only check whether the processes it posits generate outputs of the same mental state type.[9]

Of course, identifying dual process theories isn't the only purpose for which we might want to determine whether two tasks are of the same type. Suppose, for example, that we are interested – as we will be in Chapter 2 – in whether the processes in a dual process pair can generate inconsistent commitments (e.g. a commitment to *p* and a commitment to ~*p*). Here, again, we want to know whether the outputs of the processes are of the same type. However, the standard for sameness of type of outputs is higher in this context than in the previous one. For this purpose, not only mental state type similarity but also subject matter and domain similarity are relevant. When we ask whether two processes perform the same type of task in this context, we are asking not only whether they generate mental states of the same type (e.g. beliefs), but also whether they are about the same subject matter (e.g. proposition *p*) and in the same domain (e.g. theoretical). After all, mental states that differ in any of these three ways typically don't entail inconsistent commitments. There isn't a conflict between a belief that *p* and a desire that ~*p*, for example. Nor is there inconsistency between beliefs with different subject matters or in different domains; a belief that *p* doesn't conflict with a belief that ~*q*, and a moral belief that *p* doesn't conflict with a practical belief that ~*p*. To get genuinely inconsistent commitments, we need mental states that are of the same type, about the same subject matter, and in the same domain.

So, generally speaking, two tasks are of the same type provided that they satisfy the teleological criterion. Exactly how we apply the teleological criterion – and how

[9] Of course, pairs of tasks that are also similar along one or both of the other dimensions will tend to be more *interesting* than tasks that are only similar in the first sense. However, these further dimensions of similarity aren't strictly necessary for a theory to qualify as a dual process theory.

demanding the standard of sameness of type of task is – depends on the purpose for which we are trying to identify tasks of the same type. If we are just trying to determine whether a theory should be classified as a dual process theory, the standard is not very high; the outputs of the processes posited by the theory need only be of the same mental state type. If we want to know whether two processes can generate inconsistent commitments, on the other hand, the standard is higher; the processes' outputs must be of the same type, about the same subject matter, and in the same domain.

1.3 Different Ways of Performing the Same Type of Task

According to the second part of the dual process hypothesis, the processes in dual process pairs perform the same types of tasks in different ways. There are two possible kinds of relationships between processes that perform tasks in different ways:

(1) Converging

(2) Diverging

Converging processes are processes that perform tasks in different ways, but consistently produce the same outputs. Examples of converging processes are using long division to divide numbers and using a calculator to divide them. Assuming competence with both long division and calculator use, employing either of these processes will generate the same output. Diverging processes are processes that perform tasks in different ways, and *don't* (necessarily) produce the same outputs. Unlike the outputs of converging processes, the outputs of diverging processes can differ from, and even conflict with, each other. Examples of diverging processes are using your children's birthdays to pick lottery numbers and using a random number generator to pick the numbers. Though these two

processes might generate similar outputs by coincidence, they could (and, most likely, *would*) produce different outputs.

What makes one pair of processes converging and another diverging? Why do converging processes generate outputs that converge while diverging processes generate outputs that (at least potentially) diverge? Converging processes employ essentially the same mechanisms as each other while diverging processes use different mechanisms. The differences between converging processes are superficial differences in the way the same basic operation is implemented or executed. Long division and calculator-aided division both perform the same operation; they just use different tools to do it. The differences between diverging processes, on the other hand, are differences in the actual operations performed. Drawing on your children's birthdays and using a random number generator are deeply different ways of picking lottery numbers; the operations performed in each case are fundamentally different.[10]

Though many dual process theories (particularly in the early days of dual process theorizing) developed independently of each other, dual process theorists have tended to settle on similar characterizations of the processes. As suggested by the examples in §1.1, one of the processes in a dual process pair is typically characterized as conscious while the other is unconscious. Conscious processes are also taken to share all or a subset of other properties. Among these properties are being rule-based and reason-driven. Unconscious processes are thought to share all or a subset of a different property cluster.

[10] Strictly speaking, we should add a qualification here. Most pairs of processes that employ fundamentally different mechanisms will be diverging. However, there are some possible exceptions. Suppose that two processes employ different mechanisms, but these mechanisms are 'yoked together' in such a way that they consistently produce converging outputs. Though these processes would employ different mechanisms, they would not be diverging processes. I don't include this qualification in my main discussion of converging and diverging processes because the mechanisms employed by the processes in the dual process pairs I discuss don't seem to be yoked together in this way.

For example, they are typically described as associative and / or emotion-driven.[11] These kinds of property differences are *operational* differences; processes that are reason-driven will draw on different features of cases when making decisions, forming beliefs, etc. than processes that are emotion-driven. At least as they are characterized by dual process theorists, then, the differences between dual processes seem to be fundamental differences in the operations performed rather than superficial differences in execution of the same basic operation. As characterized by dual process theorists, the relationship between dual processes is *diverging* rather than *converging*.

Why do dual process theorists characterize the processes in these ways? More importantly, why think they're *right* to do so? One answer is that the processes in dual process pairs tend to activate different neural regions, which have been linked to different cognitive mechanisms. Some evidence for this comes from neuroimaging studies. For example, consider some recent neuroeconomics work with ultimatum games. In an ultimatum game, participants are paired with a partner. The partner is given a small sum of money – say, $10 – and told to split it with the participant however he would like. The participant is then given a choice between accepting the proposed split and rejecting it. If she accepts the split, she and the partner each get the percentage proposed but, if she rejects it, neither of them gets anything. The economically rational response to all offers of more than $0 is to accept. After all, accepting the offer nets the participant some money, rejecting it nets her nothing, and something is always better (economically speaking) than nothing. Interestingly, however, participants in ultimatum games *don't*

[11] The properties I list here are not the only properties dual process theorists attribute to conscious or unconscious processes. Conscious processes are often also described as evolutionarily recent, slow, low-capacity, etc. while unconscious processes are described as evolutionarily old, fast, high-capacity, etc. I don't discuss these properties here because the current question is whether there are mechanistic differences between the processes, and these properties don't directly speak to this question.

accept all non-zero offers. American participants tend to reject offers of less than about a third of the total sum (Güth et al., 1982).[12]

Given its economic irrationality, rejecting non-zero ultimatum game offers is a puzzling behavior. In an effort to explain it, Sanfey et al. (2003) asked participants to play the game in a functional magnetic resonance imaging (fMRI) scanner. They found that low offers consistently activated three specific neural regions: dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC), and anterior insula (AI). When activations in DLPFC outstripped activations in AI, participants tended to accept the low offers. When AI activations were stronger than DLPFC activations, on the other hand, they tended to reject them. DLPFC has been linked to cognitive functions like goal maintenance and executive control, AI is implicated in emotional (particularly disgust) processing, and ACC is involved in the resolution of cognitive conflict. Therefore, this pattern of neural activations suggests the following picture of economic decision-making: there are two processes – one a 'cold,' reason-driven process and the other a 'hot,' emotion-driven process – that compete to generate responses to ultimatum game offers.

Greene et al. (2001) tell a similar story about moral reasoning. They asked participants to make a judgment about the following trolley case while in an fMRI scanner:

> A runaway trolley is headed for five people who will be killed if it proceeds on its present course. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Ought you to save the five others by pushing this stranger to his death? (Greene et al, 2001, p. 2105).

[12] I specify that this is the American response because responses to ultimatum games vary cross-culturally (Oosterbeek et al., 2004). Perhaps the most surprising finding is that members of the Au and Gnau societies in Papua New Guinea reject not only very low but also very high (over 50%) offers (Henrich et al., 2001).

Participants' neural responses to this case were similar to the responses to the ultimatum game cases. Thinking about this case preferentially activated DLPFC, ACC, and medial prefrontal cortex (MPFC). As noted above, DLPFC is associated with goal maintenance, and ACC is associated with cognitive conflict resolution. MPFC, like AI, is implicated in emotional processing. Like economic decision-making, then, forming a moral judgment seems to involve competition between two processes. More specifically, it seems to involve competition between (roughly-speaking) reason- and emotion-driven processes.[13]

Further evidence that the processes in dual process pairs tend to activate different neural regions comes from studies with patients with brain damage. Patients with damage in the fusiform gyrus tend to display selective deficits in explicit visual facial recognition; they are prosopagnosic. Though they can identify other objects perfectly well, and can come to recognize people via other sense modalities (e.g. tactilely), patients with prosopagnosia are selectively impaired at visually recognizing faces. They aren't able to recognize friends, family, or even themselves in mirrors. Despite this, however, they display evidence of *implicit* facial recognition. For example, Bauer (1984) showed a patient with prosopagnosia two sets of faces: one of famous people and the other of loved ones. The patient couldn't spontaneously identify any of the faces, and performed at chance when allowed to select from a multiple choice list of names for each face. However, skin conductance measures revealed that he implicitly recognized the faces; his skin conductance responses accurately discriminated correct face-name pairs from

[13] Further evidence that the processes in dual process pairs activate different neural regions is provided by Goel & Dolan (2003) and Satpute & Lieberman (2006). In some of the earliest work on dual processes in reasoning, Evans et al. (1983) showed that we use two different processes to make judgments about the validity of syllogisms. Goel & Dolan demonstrated that these two processes activate different neural regions. Dual process theories are also very popular in social cognition. In an ambitious project, Satpute & Lieberman posit two separate neural systems – which they label the X- and C-Systems – underlying these social cognition dual processes.

incorrect pairs. Similar findings with other implicit measures (e.g. speed of face matching) suggest the same conclusion (De Haan et al., 1987). Like Bauer's finding, these findings suggest that selective impairment of explicit face recognition can be accompanied by selective sparing of implicit face recognition. This suggests that explicit face recognition implicates the fusiform gyrus while implicit face recognition does not.[14]

The neural activation findings provide indirect evidence that dual processes can produce diverging outputs. There's also more direct evidence: many dual processes actually *do* produce diverging outputs. Some of this evidence is implicit in the studies described in §1.3 and at the beginning of §1.1. For example, the two processes posited by Sanfey et al. (2003) generate different responses to the same ultimatum game offer, and Greene et al.'s (2001) processes generate different responses to the same trolley case. Similarly, the mindreading processes suggested by Clements & Perner's (1994) and Ruffman et al.'s (2001) work generate different mental state attributions to the same agent in the same situation.

This kind of pattern can also be observed in many other areas of human cognition. It's evident in our evaluative, doxastic, and interpretive responses to stimuli (attitudes, beliefs, and analyses). It also recurs in our recollections of past events and the considerations that drive our thoughts and behaviors (memories and motives). Let's look at some examples of each of these kinds of conflicts, starting with evaluative conflicts. Imagine you are asked how you feel about African-Americans. Do you harbor racist attitudes? Do you consider African-Americans inferior to European-Americans?

[14] For more examples of double dissociations between related conscious and unconscious capacities, see work with patients with amnesia (Graf & Schacter, 1985; Schacter & Graf, 1986), blindsight (Weiskrantz, 1986; Weiskrantz et al., 1974), dyslexia (Shallice & Saffran, 1986; Coslett, 1986), and aphasia (Blumstein et al., 1982).

Presumably, your answers to these questions will be 'no'; when explicitly asked about your racial attitudes, you'll probably deny any racism. Now, your denial *might* be disingenuous. Because we live in a time and place where racism is (in most circles) considered unacceptable, people who know they are racist sometimes disavow negative attitudes toward other races to avoid social censure. For most readers of this work, however, the denial is *not* disingenuous. You genuinely take yourself not to be racist and your self-reports to be accurate expressions of your racial attitudes. Not only do you *assert* that you don't have racist attitudes, but you genuinely (explicitly) *believe* that you don't have racist attitudes.

Now suppose that you are asked to complete a racial implicit attitude test (IAT). In an IAT, participants are presented with five sets of stimuli and two categories, and asked to sort the stimuli into the categories as quickly as possible. The speed with which participants sort the stimuli reveals the implicit associations they draw between the categories. In a standard race IAT, the categories are racial (African-American vs. European-American) and valenced (good vs. bad), and the stimuli are pictures of African- and European-Americans and valenced words. If you're like most participants – even most participants who are (non-disingenuously) explicitly committed to egalitarianism – your performance on the race IAT will reveal a moderate to strong implicit association between 'African-American' and 'bad' (Greenwald et al., 1998). If you're like most participants, there is also little correlation between this implicit attitude and your explicit racial attitudes (as gauged by measures like the feeling thermometer, Modern Racism Scale, and Discrimination Scale). Participants' implicit and explicit racial attitudes tend to diverge quite drastically (Greenwald & Banaji, 1995; Greenwald et al., 1998). And,

interestingly, these divergent impressions of the same token stimulus (e.g. a given human being) can be held simultaneously (Uleman et al., 2005).[15]

Like attitudes, consciously- and unconsciously-produced beliefs can dissociate. The mindreading studies cited in §1.1 provide some confirmation of this. Further evidence for it comes from work with Pavlovian, or associative, conditioning. Perruchet (1985) used a 50% random partial reinforcement schedule to condition participants to associate tones with puffs of air. On this kind of schedule, the actual probability that there will be an air puff on any given trial is 50%. Between trials, Perruchet used conscious and unconscious measures to test participants' expectations that there would be an air puff on an upcoming trial.

Surprisingly, participants' responses on these tests didn't line up with the actual probability that there would be an air puff; participants didn't expect air puffs to occur at the 50% rate. Even more surprisingly, their conscious and unconscious expectations diverged. When explicitly asked whether they thought there would be an air puff on the next trial (conscious measure), they displayed the gambler's fallacy; they were more likely to expect a puff on the next trial following a string of no-puff trials, and less likely to expect a puff after a string of puff trials. Unconscious measures (of eye blink responses), on the other hand, told the opposite story. According to these measures, participants displayed the opposite of the gambler's fallacy; they were more likely to expect a puff if there had been puffs the previous few trials, and less likely to expect one if there hadn't been puffs. Another way to phrase these findings is as follows: after a

[15] These kinds of results are not limited to racial attitudes. There's evidence for divergence in implicit and explicit attitudes toward women (Banaji & Greenwald, 1995) and lesbians and gay men (Steffens, 2005). Harvard University's Project Implicit is also conducting ongoing research into implicit attitudes toward a range of other minority groups (e.g. the elderly, the disabled, Native Americans, etc.) and activities (e.g. smoking).

string of no-puff trials, participants consciously believed there would be a puff on the next trial but unconsciously believed there wouldn't be one and, after a string of puff trials, they consciously believed there wouldn't be a puff on the next trial but unconsciously believed there would be one.

Similar patterns crop up with linguistic analysis. In a 1984 study, Groeger primed participants with a word then asked them to complete a sentence with either of two target words: a word that was semantically related to the prime or a structurally-related word. For example, he primed participants with the word 'snug' then asked them to complete the sentence, 'She looked […] in her fur coat' with either 'cozy' or 'smug.' In one condition, the prime was presented subliminally (unconscious condition) and, in the other, it was presented above the audible lumin (conscious condition). Groeger found that participants in the conscious condition were more likely to complete the sentence with the *structurally*-related target word (e.g. 'smug') while unconscious condition participants were more likely to choose the *semantically*-related option (e.g. 'cozy').

Divergence isn't unique to processing of currently present stimuli. Consciously- and unconsciously-produced memories can also diverge. Strack & Deutsch (2004) found that the processes in social decision-making dual process pairs process negations differently. For example, given the statement, 'Sam is not messy,' the conscious process processes 'not messy' as a single property while the unconscious process processes 'not' and 'messy' separately. Drawing on this finding, DeCoster et al. (2006) developed a paradigm to test whether the processes in memory dual process pairs can generate conflicting memories about the same stimuli. They presented participants with pictures paired with trait information, and asked them to copy the information into a booklet.

Some of the trait information was about a trait the pictured individual possessed (e.g. Phil is smart) while other information was about a trait the individual lacked (e.g. Sam is not messy). After a short distracter task, DeCoster et al. tested participants' memories for the pairings. Some participants were given an explicit memory test while others took an implicit test. DeCoster et al. found that participants in the former group tended to remember the correct associations between the pictures and the negated traits. Participants in the latter group, on the other hand, tended to associate the pictures with *non*-negated versions of the traits. For example, explicit test participants remembered a link between Sam and 'not messy' while implicit test participants associated Sam with 'messy.'

These findings led DeCoster et al. (2006) to posit the existence of two neuroanatomically distinct memory systems:

(1) An unconscious slow-learning system

(2) A conscious fast-binding system

As the above-described findings show, the products of these two systems are not necessarily correlated with each other. DeCoster et al. explain that,

> Even though the content of both memory systems is ultimately shaped by the same experiences, differences in the way that the two systems process this information could potentially lead to different representations of the event. The fact that the two systems store their representations in different areas of the brain means that any inconsistencies between them don't have to be resolved (2006, p. 9).

This general conclusion – that there are two memory systems whose outputs needn't line up with each other – has also been confirmed in other studies (see, for example, Kunst-Wilson & Zajonc, 1980).

Finally, conscious and unconscious motivations can dissociate. Using the Thematic Apperception Test (TAT) and self-report as measures of implicit and explicit motives, respectively, McClelland and colleagues (McClelland et al., 1953; McClelland et al., 1989) showed that implicit and explicit motives can diverge. In a typical TAT, participants are shown ambiguous pictures, and asked to construct a narrative for them. Their responses are then coded for a variety of motivational dispositions, like achievement, association, and power. The motives that emerge from responses to TATs show little correlation with self-reported motivational dispositions (McClelland, 1980). The fact that a participant displays a significant implicit disposition to be motivated by power, for example, doesn't necessarily tell us anything about his explicit power motive disposition. Of course, he might be explicitly motivated by power. However, he equally might not. His conscious and unconscious propensities to be driven by power can diverge from each other.

Now, there's an objection that might be lodged against this kind of conclusion. According to this objection, the reason TAT- and explicitly-measured motives differ isn't that conscious and unconscious motives diverge. Rather, it's that the TAT isn't really measuring *motives* at all. Like the IAT (and, indeed, many implicit measures), the TAT has its detractors. These detractors are skeptical that the TAT is really tapping into the psychological phenomena its proponents *claim* it's tapping into. They doubt that the TAT is really measuring motives.

However, though skepticism might seem natural here, it's not really warranted. The TAT and explicit measures of motives predict different behaviors. TAT measures are better at predicting long-term, spontaneous behavioral trends while explicit measures are

more closely connected to more immediate and specific behavioral responses. For example, conscious measures of affiliation motives are better predictors of group project choices while unconscious measures are better at predicting the likelihood that an individual will be engaged in conversation when randomly beeped (McClelland, 1980). The same is true of the IAT and explicit attitude measures. For example, conscious measures of racial attitudes tend to be better predictors of assessments of the guilt of an African-American defendant while unconscious measures are better at predicting friendliness toward an African-American experimenter (Wilson et al., 2000). The TAT's and IAT's behavioral prediction successes suggest that they are tapping into real parts of participants' motivational and evaluative structures. And the divergence between TAT- / IAT- and explicit measure-predicted behaviors suggests that conscious and unconscious motives / attitudes really do diverge.

The above is just a sampling of the relevant dual process work. However, it nicely conveys the tenor of the literature. In general, the processes in dual process pairs are described in operationally different terms; they are assigned different clusters of properties that seem to reflect the use of different sets of mechanisms. These kinds of descriptions suggest a diverging rather than a converging picture of the relationship between the processes. And this conclusion seems empirically well-motivated. For one thing, the processes in dual process pairs tend to activate different neural regions, which are linked to different cognitive mechanisms. For another, the processes can, and sometimes do, generate different outputs. Indeed, they can even generate *conflicting* outputs.[16]

[16] Drawing on the process dissociation procedure methodology pioneered by Jacoby (1991), Jacoby et al. (1997) offer a similar characterization of the relationship between conscious and unconscious processes.

1.4 Varieties of Dual Process Theories

So far, I've been emphasizing the similarities between dual process theories. However, there are also some important differences between them. We can distinguish three broad classes of dual process theories:

(1) Independent

(2) Parallel

(3) Serial

Recall that tasks are defined teleologically, or in terms of the ends toward which they are directed. A consequence of this definition is that two processes that perform the same type of task can take entirely different types of inputs. This is the case, for example, with Berry & Broadbent's (1984) learning processes. Though both of these processes perform learning tasks, one takes verbal information as an input while the other takes practice as its input. Because they respond to different inputs, these kinds of dual process pairs are independent of each other.

The inputs to processes in parallel and serial dual process pairs are more closely linked. Parallel dual processes take the same type of – and sometimes even the same *token* – input. For an example of a parallel dual process pair, think about the mindreading processes described in §1.1. Both of those processes take information about mindreading targets as their inputs. Other examples of parallel dual processes are Sanfey et al.'s (2003) economic decision-making processes and Greene et al.'s (2001) moral judgment processes. Sanfey et al.'s processes both respond to economic choices, and Greene et al.'s processes both respond to moral dilemmas. Because they can respond to the same token

---

They say that "conscious and unconscious influences act fully independently of each other. Conscious processing can happen without unconscious processing and vice versa" (1997, p. 19).

input, parallel dual processes can run in parallel with each other, operating on inputs and generating outputs at the same time.

The inputs to serial dual processes are linked in a different way: the inputs to one process in a serial dual process pair are consequences of the outputs of the other. To understand how this works, think about Haidt's (2001) moral justification-generation processes. The unconscious process in that pairing generates justifications, which prompt a judgment about the case under consideration. This judgment is then taken as an input to the conscious justification-generation process. Though the two processes in serial dual process pairs don't take exactly the same inputs, there's a clear connection between them; they aren't entirely disconnected in the way independent dual processes can be.

Haidt's (2001) serial dual process theory of moral reasoning is a specific application of a general approach to reasoning that was developed by Wason & Evans (1975). In the 1960s, Wason (1966) developed a task that has come to be known as the Wason selection task (WST). In a typical version of this task, participants are presented with the following four cards:

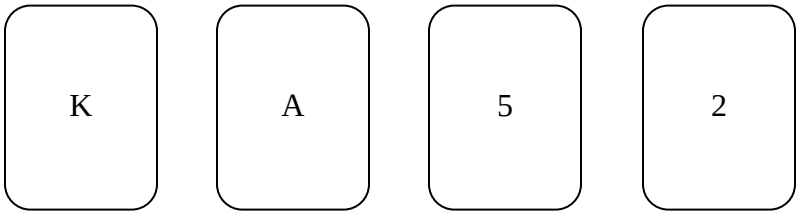

Figure 5

They are then asked to say which of the cards would have to be turned over to test the following rule: if a card has an even number on one side, it has a vowel on the other. The correct answer is that the 'K' and the '2' would have to be flipped. However, Wason found that most participants respond with the '*A*' and the '2.' His initial explanation of

this finding was that participants tend erroneously to pursue a verification strategy; instead of choosing the cards that would *falsify* the rule, they choose the ones that would *verify* it. This explanation seemed to be borne out by participants' introspective reports. When asked why they gave the answers they did, participants who chose the 'A' and the '2' tended to say that they were trying to verify the rule.

Wason & Evans (1975) offer a different explanation. They hypothesized that participants' responses to the WST were due to a matching bias. Rather than pursuing either a verificationist or a falsificationist strategy, participants were simply choosing the cards that were mentioned in – or that *matched* – the rule; the rule mentions vowels and even numbers, so participants chose the vowel card and the even number card. To test this hypothesis, Wason & Evans presented participants with two rules. The consequent of one of the rules was like the consequent of the rule described above (e.g. it has an even number on the other side). The consequent of the other rule was a doubly negated version of the consequent of the first rule (e.g. it *doesn't* have an *odd* number on the other side). If participants were employing a verificationist strategy, we would expect them to respond to both rules in the same way. If their responses were due to a matching bias, on the other hand, we would expect them to be different. Wason & Evans found that participants' responses to the two rules were different.

In addition to undermining Wason's (1966) original explanation of responses to the WST, this finding raises questions about participants' introspective reports. If participants' responses are *actually* due to a matching bias, why do they *report* that they are due to use of a verification strategy? Wason & Evans' (1975) answer is that – as in the moral reasoning case – there are two processes at work here: a conscious process and

an unconscious process. Typically, the unconscious process produces the justifications that prompt the actual response to the WST while the conscious process produces a *post hoc* justification of this response. Like Haidt's (2001) theory of the processes that generate moral justifications, Wason & Evans' explanation of WST performance is a serial dual process theory.

An interesting feature of serial dual processes like those posited by Haidt (2001) and Wason & Evans (1975) is that they have to operate in sequence. Unlike independent and parallel dual processes, which can (at least in principle) operate at the same time, the processes in serial dual process pairs have to run serially. This is a consequence of the way they work and, more specifically, the types of inputs they take. Because the conscious processes in serial dual process pairs take consequences of the outputs of the unconscious processes as inputs, the conscious processes can't run until *after* the unconscious processes are complete.

Independent dual processes won't loom particularly large in the upcoming discussion. Though they have interesting practical implications (e.g. for education policy), they don't have the same kinds of metaphysical implications parallel and serial dual processes do. My focus in Chapter 2 is, therefore, on parallel and serial dual processes. I use discoveries about these kinds of dual processes to challenge a common philosophical assumption about the mind. Philosophers tend to assume that there is one person in each normal adult human body. In Chapter 2, I argue that this assumption is mistaken. Given parallel and serial dual process findings like those described in this chapter, the average normal human being is analogous to a pair or group of human

beings. Like pairs and groups, the average normal human being doesn’t contain a single, unified person.

CHAPTER 2: DIVIDED MINDS

2.1 Odd Couples

Imagine two human beings – call them 'Felix' and 'Oscar' – who have deeply different ways of engaging with the world. Felix carefully calculates his responses to things while Oscar is impulsive. Felix cares deeply about advancing at work while Oscar is more concerned with having a good time. Felix considers cleanliness next to godliness while Oscar doubts the state of his apartment has any bearing on his prospects for eternal salvation. If you were to encounter Felix and Oscar, you'd presumably have little hesitation about describing them as not-the-same-person.

Now imagine that, tragically enough, Oscar passes away. Shortly thereafter, Felix starts behaving strangely. He puts off cleaning well beyond the point at which he would previously have felt driven to it. He procrastinates on writing his weekly news column until mere hours before the deadline. He starts making emotion-driven decisions, abandoning his customary detailed pro-and-con lists. He also experiences internal conflicts about his thoughts and behaviors. Whereas his beliefs, behaviors, etc. had previously seemed to flow seamlessly from his reasoning and intentions, there sometimes now seems to be more of a battle over what he thinks, says, and does.

Troubled by these shifts, Felix decides to check in with his doctor. The doctor, at a loss, refers him to the local religious authority. There, Felix gets the bad news: he has been inhabited by a *dybbuk*. According to Jewish lore, the religious authority explains, a *dybbuk* is the dispossessed soul of a deceased individual – in this case, apparently Oscar – who wasn't able to accomplish all he wanted to during his lifetime. After death, the *dybbuk* inhabits someone who's still alive to try to achieve his goals. Felix's newfound

slovenliness, procrastination, and emotionality is Oscar's attempt to finish unfinished business. There's still some Felix in Felix, but it now has to compete with *dybbuk*-Oscar to determine which beliefs, desires, etc. the human being they share will have and which behaviors it will perform.

Here, again, we don't seem to have a single person. Like Felix and Oscar, Felix and *dybbuk*-Oscar have deeply different ways of thinking about the world. They are driven by different motivations, and swayed by different desires. They approach problem-solving in different ways, and are committed to different beliefs. Felix and *dybbuk*-Oscar have different *personalities*, and these different personalities compete for control over the body that contains them. If we watch Felix for a few days, we'll see that he sometimes acts in a Felix-y way and sometimes in an Oscar-y way. When the Felix personality wins out, he cleans up the kitchen immediately after making dinner while, when the Oscar personality wins out, he leaves empty pizza boxes scattered around the apartment. A natural explanation of these conflicting behaviors is that there are two different forces inside Felix's body that are competing for expression. Felix now has to contend with another force (*dybbuk*-Oscar) to determine what Felix's body will think, say, and do.

We tend to think of ourselves as pre-*dybbuk* Felixes. According to the parallel dual process findings reported in Chapter 1, however, we're actually more like *post-dybbuk* Felix. Rather than a single, unified front, those findings suggest, we house two (or more) forces with different personalities that compete for control over our thoughts and behaviors.[17] The processes in our dual process pairs use different mechanisms to address

[17] Some dual process theorists suggest that there are two systems in the mind: one composed of the conscious processes in dual process pairs and the other composed of the unconscious processes. In later

the same types of questions, so they can (and often do) come up with different answers. They generate different motivations, beliefs, and desires, and these different mental commitments incline us to different behaviors. The processes battle each other to determine which of these commitments and behaviors are expressed by our bodies as wholes.

If we look closely, we can sometimes even see these battles playing out. For example, this morning, I had a strong conscious desire to leap out of bed, into my desk chair, and right to work on this chapter. However, the *dybbuk* had other ideas. Rather than feverishly writing, I found myself curled in a comfy chair with coffee and the most recent issue of *The Economist*. Similarly, influenced by my utilitarian-leaning colleagues at Rutgers, I tend to endorse the consequentialist judgment about the footbridge version of the trolley case. Nonetheless, I find it hard to shake the feeling that there's something wrong with this judgment, and that I really have no moral business pushing a man in front of a runaway trolley. This feeling is the *dybbuk* niggling at conscious me, trying to make itself heard.

It's important to distinguish these deep, troubling conflicts from the unproblematic, run-of-the-mill conflicts we experience every day. For an example of a run-of-the-mill conflict, imagine a case in which I am trying to decide whether to spend the evening revising my dissertation or re-watching Season 1 of *The Wire*. I draw up a pro-and-con list for each option, weigh the pros against the cons, and arrive at a decision. In this case, though there are competing considerations in play, they are all entertained in the context of a single process. I employ one decision-making process, which happens to

---

chapters, I assess this proposal. For now, though, I don't assume that there are *exactly* two forces at work. I just claim that there's more than one.

draw on multiple different considerations. Contrast this with the dual process cases. In those cases, there are two different processes occurring simultaneously. The run-of-the-mill case – which involves only a single process – is naturally thought of in terms of a single subject; there's a single subject, using a single process to arrive at a decision. Such a reading of the dual process cases isn't nearly as natural. Those cases are more naturally described as involving two subjects – one employing each of two different processes.

If this is right, human beings are not single, unified persons. The combination of Felix and *dybbuk*-Oscar (Felix / *dybbuk*-Oscar) seems not to be a single, unified person so, if normal human beings are like Felix / *dybbuk*-Oscar, they're not single, unified persons either. Now, this conclusion is bound to be controversial. For one thing, we have a strong intuition that we *are* single, unified persons. For another, denying this intuition could cause us all sorts of trouble. For example, it would force us to rethink the way we parcel out moral responsibility. So, there's strong incentive to deny the current conclusion; there's strong incentive to say that the appearance that we're like Felix / *dybbuk*-Oscar and that Felix / *dybbuk*-Oscar isn't a single, unified person is misleading. At some point, we want to say, the current argument goes awry.

Of course, though, if we want to reject its conclusion, we have to pinpoint the place at which the argument goes off-track. There are two steps to the argument:

(1) Felix / *dybbuk*-Oscar is not a single, unified person.

(2) Normal human beings are like Felix / *dybbuk*-Oscar (in personhood-relevant ways).

To avoid the conclusion that normal human beings are not single, unified persons, we have to deny one (or both) of these premises. Is either of them assailable? In §2.2, I consider challenges to the second premise and, in §2.3, I entertain objections to the first. I

close, in §2.4, with a summary of the chapter's conclusions and their implications for analyses of personhood.

2.2 The Analogy between Felix / *Dybbuk*-Oscar and Normal Human Beings

According to the second premise of the current argument, normal human beings are like Felix / *dybbuk*-Oscar in personhood-relevant ways; the conclusions we draw about the personhood status of Felix / *dybbuk*-Oscar also extend to normal human beings. To challenge this premise, we would have to sever the link between Felix / *dybbuk*-Oscar and normal human beings. The dialectical strategy here is to find personhood-relevant differences between normal human beings and Felix / *dybbuk*-Oscar that render them disanalogous. There are certainly *some* differences between the two cases. Let's see whether any of them break the analogy I've drawn between them.

One obvious difference between the two cases is that normal human beings are entirely composed of human matter while Felix / *dybbuk*-Oscar is half-spirit matter. Because he is a spirit, *dybbuk*-Oscar is an obviously alien force. This, it might be argued, is what prevents him from being integrated into Felix, and Felix / *dybbuk*-Oscar from being a single, unified person. The normal human being faces no such obstacle. Both forces in the normal human being are made of human matter. Therefore, neither is obviously alien to the other in the way *dybbuk*-Oscar is to (human) Felix.

To test whether this is a relevant disanalogy between the cases, imagine that the interloper into Felix's psyche is not *dybbuk*-Oscar, but Oscar himself. Felix and Oscar are strolling through a park, having a nice chat, when they stumble into a Star Trek-style teletransporter. The teletransporter scrambles Felix and Oscar physically, but keeps them largely separate psychologically. The end result is that they are trapped inside the same

body, but retain their separate, distinctive personalities. Much like Felix and *dybbuk*-Oscar, the being that steps out of the teletransporter contains two different personalities that struggle against each other to control the body they share. This case, it seems, differs very little from the Felix / *dybbuk*-Oscar case. Here, as there, the most natural explanation of the behaviors we would observe is that there are two persons trapped inside the same body. If asked to provide a description of what we saw, we'd talk in terms of Felix competing with Oscar, Felix winning out in one case, Oscar winning out in the other, etc. The alienness of *dybbuk*-Oscar does not seem, therefore, to be playing an important role in our judgments about Felix / *dybbuk*-Oscar.

Of course, it might be argued that the teletransporter introduces an alien element of its own. After all, outside the realm of sci-fi, teletransporters don't exist. Wilkes (1988) has argued that outlandish sci-fi thought experiments are poor gauges of our opinions about philosophical issues. They involve too many uncontrolled variables – a world in which teletransporters existed would differ from our world in more ways than just this one – for us to draw any definitive conclusions from our intuitions about them.

I tend to doubt that the kinds of differences that would be required for teletransporters to exist would affect our intuitions about teletransporter-scrambled Felix and Oscar. Fortunately, though, the general point here doesn't depend on an appeal to teletransporter-scrambled Felix and Oscar. They have real-life analogs to which we can appeal: dissociative identity disorder patients. As the original label for the disorder (multiple personality disorder) suggests, patients with dissociative identity disorder contain multiple personalities. When different personalities, or 'alters,' assume control of the patient's body, he displays different preferences, holds different beliefs, is driven by

different motives, etc. For example, one alter might be honor-driven or responsible in a way another isn't. When the former alter is in charge, perceived slights are significantly more likely to lead to bar fights than when the latter is in control. Similarly, the former alter is a better bet for entrusting with important projects than the latter.[18] Though the relationships between alters can vary from patient to patient (and even alter to alter), alters often compete with each other for control over the patient as a whole.

As in the scrambled-teletransporter and Felix / *dybbuk*-Oscar cases, the most natural way to describe the dissociative identity disorder case is in terms of a variety of different forces competing for control over a single human body. There are deep psychological schisms in the dissociative identity disorder patient which aren't compatible with thinking of him as a single, unified person. His thoughts and behaviors are inconsistent from one moment (when he is controlled by one alter) to another (when control transfers to another alter). This kind of inconsistency makes sense if we think of him as containing multiple persons in a way that it doesn't if we try to conceive of him as a single, unified person. Importantly, the dissociative identity disorder patient is composed entirely of human matter; there are no *dybbuk*-like spirits involved. This kind of case confirms, then, that our reactions to the Felix / *dybbuk*-Oscar case aren't tapping into the alienness of *dybbuk*-Oscar.

So, the fact that Felix / *dybbuk*-Oscar is not wholly human and the normal human being is doesn't seem to mark a relevant difference between the two cases. Of course, though, that isn't the only difference between them. Here's another. By hypothesis, Felix and *dybbuk*-Oscar each start out as a complete, fully-realized person. The same isn't

[18] For a more thorough, and colorful, account of a dissociative identity disorder patient, see Prince's (1909) account of Christine Beauchamp.

obviously true of the processes in dual process pairs. So far, we only know that there are pairs of processes that perform the same types of cognitive tasks in different ways. As of yet, we haven't established that these processes resolve into separate unified beings. As of yet, we don't have reason to believe that the conflicts between processes in dual process pairs are conflicts between *persons*.

Now, one possible response to this objection is that the processes in dual process pairs *do* resolve into separate unified beings, and the conflicts between them *are* conflicts between persons. In Chapters 3-5, I offer some support for this response. There, I argue that the set of unconscious states and processes can perform the functions we expect of persons, and is unified in the way we expect persons to be.

However, we don't need to accept this argument – or think the unconscious is a person – to have a response to the current objection. The idea behind the current objection is that an opposing force within a being only challenges the single, unified personhood of the being if it is, itself, a single, unified person. But why think this is the case? Imagine that I create a nanorobot that performs only one function – say, compelling you to order sushi every Monday at 7pm – and implant it into your brain. You retain control over yourself most of the time but, on Mondays at 7pm, the nanorobot takes over. In this case, the nanorobot clearly isn't a complete, fully-realized person. However, it also isn't part of you. The combination of you and the nanorobot is not a single, unified person. Rather, it's a person plus a nanorobot. This suggests that something needn't itself be a single, unified person to challenge something else's single, unified personhood.

Another (alleged) difference between Felix / *dybbuk*-Oscar and normal human beings is a phenomenological difference. Recall from my initial story about post-*dybbuk*

Felix that he learned about the *dybbuk* because he noticed himself engaging in strange patterns of behaviors, and felt internal conflicts that he hadn't previously experienced. Post-*dybbuk* Felix *felt* as if he were occupied by an opposing force; he no longer felt as if he were a single, unified person. Normal human beings, on the other hand, *do* tend to feel like single, unified persons. If someone were to ask you how many people there were inside you, you'd probably give them a strange look. This is an odd question to ask a normal human being because the answer seems so obviously to be *one*. So, there seems to be a difference in phenomenology between Felix / *dybbuk*-Oscar and normal human beings: Felix / *dybbuk*-Oscar feels like two persons while normal human beings don't.

One way to interpret this objection is as the claim that a phenomenological feeling of unity is sufficient for personhood. However, this is not a particularly charitable interpretation. After all, we don't typically think our feelings about things shape reality in this way. A better way to interpret the objection is as the claim that the phenomenological feeling taps into reality: I *feel* like a single, unified person (and Felix / *dybbuk*-Oscar doesn't) because I *am* a single, unified person (and Felix / *dybbuk*-Oscar is not).

A first thing to note about this is that there isn't quite as clear a distinction between Felix / *dybbuk*-Oscar's phenomenology and normal human beings' phenomenology as the above description suggests. Think back to some of the cases I introduced in §2.1. There, I described a case in which I consciously intend to get right to work in the morning but find myself reading *The Economist* instead and a case in which I have qualms about my consequentialist response to the footbridge version of the trolley problem. In those cases, I'm displaying the kinds of strange behaviors, and experiencing the sense of inner turmoil, that raised post-*dybbuk* Felix's suspicions. So, even normal

human beings have experiences that could lead them to doubt their single, unified personhood.

Also, the second interpretation of the current objection turns out not to be much more charitable than the first. As is now a commonplace, the way something *feels* is not necessarily indicative of the way it actually *is*. The Earth feels flat, but it's not. The table in front of me feels solid, but it's not. I feel as if I base my moral judgments on a conscious weighing of reasons, but (often) I don't. Phenomenological data *are* data, and they should be weighed in an overall accounting of evidence. However, we shouldn't let them get us carried away. They are highly defeasible, and can be overridden by conflicting evidence.

And, in the current case, there's good reason to think that the phenomenological data *are* overridden by conflicting evidence. We have compelling empirical evidence that normal human beings are not as unified as we seem to be. The extensive dual process literature clearly establishes that there are separate processes that compete to perform many of our cognitive tasks.

There's also a ready explanation for why we feel unified even though we aren't – namely, that we have a particularly powerful capacity for confabulation. This is illustrated strikingly by the serial dual process cases. In those cases, participants feel as if there is a direct line from a dilemma through their justifications for a judgment about the dilemma to the judgment itself. However, experimentation has shown that this often isn't how things really work. Instead, an unconscious process outputs a judgment, and a

separate, conscious process generates a *post hoc* justification of the judgment. We confabulate a throughline where one doesn't actually exist.[19]

Some interesting confirmation that feelings of unity don't map onto the reality of unity is provided by dissociative identity disorder patients. It's often the case that different alters have different levels of access to the operations of other alters. Some might be able to 'watch' others controlling the body while others have little or no access to what happens when they aren't in charge. In this case, different alters have different feelings of unity. The alters that have access to the others' operations are like Felix / *dybbuk*-Oscar. They're aware that there's an opposing force in the body, and that this force competes with them for control. These alters don't have a feeling of unity. The alters that don't have access to the others' operations, on the other hand, are more like normal human beings. Though they might sometimes get behavioral evidence – or a hunch – that they aren't in complete command, they are less directly aware of the presence of competing forces. If we took phenomenology to map onto reality, we'd have to conclude that dissociative identity disorder patients both are and aren't single, unified persons. Of course, though, this is impossible. This highlights the tenuousness of the link between feelings of unity and the reality of unity.

The objections I've been considering so far are based on differences between the natures of the two forces involved in the disputes. The first objection was based on a difference between the compositions of the forces: Felix / *dybbuk*-Oscar is part-spirit while the normal human being is entirely human. The second was based on a (alleged) difference in their personhood statuses: Felix and *dybbuk*-Oscar both started out as

[19] This capacity for confabulation also explains why, despite the pervasiveness of dual processes, we rarely doubt our own unity.

complete, fully-realized persons while the conflicting forces in normal human beings are (allegedly) not all persons. The third was based on a (alleged) phenomenological difference: Felix and *dybbuk*-Oscar *feel* like two different persons while normal human beings (allegedly) don't.

Differences in the natures of the forces are not, however, the only possible differences between the two cases. Another is a difference in the nature of the *conflict* between the forces. More specifically, it might be objected that there isn't a *genuine* conflict between the forces in the normal human being at all. Felix and *dybbuk*-Oscar compete for control over Felix's thoughts and behaviors. The forces in the normal human being, on the other hand, don't come into direct conflict with each other.

One way to cash out this kind of objection is to say that the forces in the normal human being don't output genuinely conflicting mental states. There are a couple of possible versions of this objection. First, recall from Chapter 1 that two mental states genuinely conflict only if they are:

(1) Of the same mental state type

(2) In the same domain and

(3) About the same subject matter

Given this, one way to cash out the current objection is to say that the outputs of dual processes differ in one or more of these ways. Now, parallel dual processes typically do generate outputs in the same domain. For example, Greene et al.'s (2001) processes both output *moral* mental states, the mindreading dual processes both generate *theoretical* states, etc. It's less obvious, though, that they generate outputs of the same mental state type or about the same subject matter. Indeed, objections could be raised to claims to

either type of similarity. Let's take a look at each of these kinds of objections, starting with objections to the claim that dual processes generate outputs of the same mental state type.

If someone tells you that he doesn't dislike African-Americans or that he believes his sister thinks her toy is in a basket, you'll probably take him at his word; you'll attribute to him a positive (or at least neutral) attitude toward African-Americans and the belief that his sister believes her toy is in the basket. However, verbal assertion is not the only way we convey mental states. Suppose, for example, that someone was consistently significantly less friendly to African- than European-Americans. What attitude would you take him to have toward African-Americans? Presumably, you'd think he had a negative attitude toward them. Similarly, suppose that, when you ask someone where he thinks his sister thinks her toy is, he responds by looking at a box. What belief about his sister's beliefs would you attribute to him? Presumably, you'd think he believed that she believed the toy was in the box. Intuitively, then, the behaviors produced by the unconscious processes in dual process pairs *read* as evidence of the same kinds of mental states as are produced by the conscious processes. The mental states measured by the IAT seem to be attitudes, the states measured by looking times seem to be beliefs, etc.

So, what could drive the objection that they are *different* types of mental states? The IAT-measured mental states are *unconscious* states; IAT-takers aren't consciously aware that they hold these attitudes. Similarly, the looking-time measured states in the mindreading case are unconscious. It's this unconsciousness – so the current objection goes – that makes these states different from the outputs of conscious processes. The outputs of the processes in dual process pairs are of different mental state types because

conscious and *unconscious* mental states are of different types. More specifically, unconscious mental states are pale imitations of their conscious counterparts. Though an unconscious mental state might be belief-*like*, for example, it can't be a genuine *belief*.

There are two possible routes to this conclusion:

(1) Unconscious mental states are different in kind from their conscious counterparts *in virtue of their unconsciousness*.

(2) Unconscious mental states are different in kind from their conscious counterparts in virtue of something *other* than their unconsciousness.

The first claim is defended for intentional states by Searle (1992). Searle argues that intentional states are intrinsically (vs. as-if) intentional, intrinsically intentional states have aspectual shapes, and aspectual shapes can't be wholly characterized in terms of third-person properties. Therefore, intentional states can't be wholly characterized in terms of third-person properties. States that aren't potentially conscious (i.e. unconscious states), on the other hand, *can* be wholly characterized in terms of third-person properties. Intentional states must, therefore, be at least potentially conscious. Unconscious states don't make the grade.

What exactly does Searle (1992) mean here? To say that a state has as-if intentionality is to say that it is intentional in only a metaphorical sense. Though we might say that a lawn is *thirsty* or that water *wants* to roll downhill, for example, what we mean is that they are in conditions or exhibiting behaviors that would be accompanied by intentional states (i.e. the desire to drink or move in a particular direction) in us. States have intrinsic intentionality, on the other hand, if ascriptions of intentionality to them are literal. Aspectual shapes can be thought of as something like 'modes of presentation' or 'descriptions.' To say that a state has an aspectual shape is to say that it is a thought about

a thing *as* a particular thing, or *under a particular description*. The gist of Searle's argument is, therefore, that literal intentional states have to be thoughts about things under particular descriptions, thoughts about things under particular descriptions have to be characterized (at least in part) in terms of conscious properties and, therefore, literal intentional states must (at least potentially) have conscious properties.

Objections have been raised to many parts of Searle's (1992) argument. For example, Dunlop (2000) identifies an internal inconsistency in it, and Dennett (1990) famously denies that intentional systems have to have intrinsic intentionality. In my view, however, the biggest problem with it is the claim that aspectual shapes can't be wholly characterized in terms of third-person properties. The problem here is nicely spelled-out by Van Baaren (1999). Van Baaren concedes that we can't give an *exhaustive* third-person account of aspectual shape. However, he argues, we don't *have* to give an exhaustive account. There are two components of aspectual shapes:

(1) Phenomenal

(2) Strictly aspectual

The phenomenal component of an aspectual shape is the way the state seems to its possessor while the strictly aspectual component is (roughly) the content of the state, or what the state is about. The phenomenal component is the part we can't capture with a third-person account. This part is not, however, integral to intentionality. The way a mental state seems to its possessor is a property of the relationship between the state and the possessor, not a property of the state itself. If it's not a property of the state, though, it's not integral to the state. What is integral to an intentional state is its strictly aspectual

component and this, Van Baaren argues, *can* be captured by a purely third-person account.

Van Baaren's (1999) contention that the strictly aspectual components of intentional states can be captured by purely third-person accounts is borne out by the findings cited in Chapter 1. Consider Groeger (1984), for example. When Groeger's participants were unconsciously primed with a word, they completed a sentence with a target word that was *semantically* related to the prime. How can we explain this? We can make sense of it only if we suppose that the unconsciously-primed participants were thinking of the prime word not only *as a word* but as a word *with a particular meaning*. In other words, they must have been thinking of the prime word under the description, 'word with particular meaning *x*.' If explaining a behavior requires reference to the content of a mental state, we can use that behavior to characterize the strictly aspectual component of the state. Therefore, we can characterize aspectual shapes in terms of third-person properties (see Nelkin, 1993 for a similar argument).

If unconsciousness itself doesn't disqualify unconscious states from counting as beliefs, desires, etc., is there some other feature of them that does? Are unconscious mental states different in kind from conscious mental states in virtue of something other than their unconsciousness? Arguably, some of Gendler's (2008) work seems to suggest that they are. Gendler describes a study – by Rozin (1986) – in which experimenters poured sugar into two bottles in front of participants then asked them to label one of the bottles 'sugar' and the other 'sodium cyanide.' Even though participants saw the sugar being poured into the bottles, and got to choose which bottle to apply each label to, they were reluctant to eat from the cyanide-labeled bottle. The participants in this study

consciously believed that both bottles contained sugar. After all, they had seen sugar being poured into them, and applied the labels to them themselves. However, their reluctance to eat from the cyanide-labeled bottle appears to signal a different – *un*conscious – belief: the substance in the bottle is dangerous, and should not be consumed.

According to Gendler (2008), though, this appearance is deceptive. Though it superficially *appears* to be a belief, the mental state signaled by reluctance to eat from the bottle isn't *really* a belief. Rather, it's an entirely different type of mental state – an alief. Though aliefs are belief-*like*, Gendler argues, they aren't actually *beliefs*. Most importantly, unlike beliefs, aliefs aren't sensitive to evidence. No matter how much evidence participants have that the contents of the cyanide-labeled bottle are non-lethal, they still hesitate to eat them.

Gendler's (2008) analysis of this kind of case arguably gives us reason to doubt that *some* unconscious apparent beliefs actually are beliefs. Aliefs are unconscious states, and some of the states we loosely describe as unconscious beliefs might actually be aliefs instead. However, her analysis doesn't warrant the conclusion that *none* of the states we're inclined to describe as unconscious beliefs are genuine beliefs. Indeed, I don't think Gendler herself means to suggest that it does. The class of aliefs is not – nor, I think, is it intended to be – *co-extensive* with the class of unconscious apparent beliefs. Rather, it's a *subset* of that class. More specifically, it's the subset of unconscious apparent beliefs that aren't sensitive to evidence. So, the class of aliefs does not include unconscious apparent beliefs that are sensitive to evidence. And there *are* some unconscious apparent beliefs that are sensitive to evidence. Consider, for example, the

unconscious attribution in the mindreading case. If the evidence available to the mindreader were to change, his unconscious attribution would also change. The existence of aliefs does not, therefore, warrant the conclusion that no unconscious belief-like states are genuine beliefs. Though some such states might be aliefs, that doesn't mean they all are.

The intuition that states like unconscious false belief attributions are genuine beliefs is a functionalist intuition. It's driven by the idea that the unconscious mental states fill the same causal roles as their conscious counterparts.[20] It might be argued, however, that this intuition is off-base; unconscious mental states *don't* really fill the same causal roles as their conscious counterparts. There are some effects we tend to associate with beliefs, desires, etc. that apparent unconscious beliefs, desires, etc. don't have (e.g. a tendency to verbally assent to a mental state attribution when asked about it). This suggests a second way to cash out the objection that unconscious mental states differ from their conscious counterparts in virtue of something other than their unconsciousness. According to this version of the objection, unconscious states fail to qualify as genuine beliefs, desires, etc. because they fail to produce some of the effects we associate with genuine beliefs, desires, etc.

Now, the conclusion that apparent unconscious beliefs, desires, etc. aren't genuine beliefs, desires, etc. doesn't fall directly out of the observation that they don't have some of the effects we tend to associate with beliefs, desires, etc. No functionalist maintains that a state has to have *all* the causal relationships we might associate with a type of mental state to count as that kind of state. For example, we might associate picking up an

[20] Note, though, that the other major (non-eliminativist) contenders for accounts of belief – representationalism and interpretivism – can also accommodate unconscious beliefs.

umbrella with the belief that it's raining, but that doesn't mean we should automatically deny that a non-umbrella-picker-upper believes it's raining. If other causal relationships we associate with the belief (e.g. donning a raincoat) are in place, the state can still qualify as the belief that it's raining. So, the fact that unconscious states are missing one (or some) of the causal relationships we associate with beliefs, desires, etc. doesn't automatically prevent them from counting as beliefs, desires, etc. To establish *that* conclusion, the objector has to show that the consciousness-associated relationships are somehow *essential* to the mental states.

However, the claim that consciousness-associated relationships are *essential* to beliefs, desires, etc. seems to be at odds with how we often think about mental states. Suppose, for example, that you consistently (but unconsciously) sabotage your efforts to get ahead at your job. Though you aren't consciously aware of it, you do things – procrastinating on important projects, forgetting the paperwork for a major merger meeting, etc. – that prevent you from getting a promotion. If this self-sabotage were pointed out to you, how would you respond? A natural response would be to say that you didn't *want* to get the promotion. Bringing the self-sabotage to your attention *reveals* a desire not to get ahead in the job. Similar lines can also be run with other mental states. For example, if you consistently treat individuals of other races poorly, that seems to *reveal* racist beliefs. The fact that these readings are plausible suggests that consciousness isn't essential to our characterizations of the mental states.

So, it's not obvious that unconscious mental states really are different in kind from their conscious counterparts. Also, even if we assume that they are, there are other possible responses to the current objection. One of these responses is specific to the

Gendler-inspired version of the objection. Gendler (2008) thinks aliefs are a different type of mental state than beliefs. Nonetheless, she still seems to think there's something inconsistent about simultaneously believing *p* and alieving ~*p*. She notes that, if someone were to learn *p* but continue to alieve ~*p*, he would be violating a norm. So, though belief and alief might not be exactly the same type of mental state, their status seems to be sufficiently similar for them to be capable of genuine conflict.

There's also another response that applies, more generally, to all versions of the current objection. As emphasized in the Introduction, all that's required for a process to qualify as unconscious is that its *workings*, or *mechanisms*, are unconscious; the inputs to and outputs of these mechanisms can be conscious. This means that the outputs of unconscious processes needn't *always* be unconscious. Sometimes, they can be conscious. And, indeed, we've seen some examples of cases in which they *are* conscious. Take Groeger's (1984) study, for example. In that case, the output of the unconscious linguistic analysis process was conscious; participants actually used it to complete the sentence. Similarly, when unconscious reasoning processes win out in cases like the ultimatum game and trolley problem cases, participants are consciously aware of their outputs. In these kinds of cases, it's irrelevant whether unconscious states are different in kind from their conscious counterparts. After all, the outputs of both processes – both the conscious process *and* the unconscious process – in these cases are conscious.

Having addressed the objection to mental state type similarity, let's turn to the objection to subject matter similarity. It's easiest to illustrate this objection with beliefs, so let's use those as our examples. For two beliefs to contradict each other, one must have the content *p* while the other has the content ~*p*. If one is believed under the description

'belief that *p*' while the other is believed under the description 'belief that *q*,' they don't directly contradict each other (even if '*q*' can be rephrased as '~*p*'). According to the subject matter similarity-based objection, the outputs of dual processes aren't believed under the contradictory descriptions. Consider the mindreading case, for example. In that case, the outputs of the conscious and unconscious mindreading processes are naturally – and perhaps even *best* – interpreted as 'belief that the girl believes her toy is in the basket' and 'belief that the girl believes her toy is in the box.' These two beliefs don't strictly contradict each other.

However, though this objection is plausible in some cases, it's significantly less plausible in others. There's an argument to be made that the beliefs in the mindreading case don't strictly contradict each other; arguably, one of the beliefs really is the belief that the girl believes the toy is in the basket (belief that *p*) while the other is the belief that she believes it is in the box (belief that *q*). But other dual process cases can't be dispatched quite as easily. In the IAT case, for example, it's difficult to deny that the attitudes generated by the two processes are both attitudes toward African-Americans. The case for this interpretation is strengthened by the evidence that each of the attitudes manifests in behaviors that are specifically directed toward African-Americans (Wilson et al., 2000).

Another – perhaps, better – response to the subject matter similarity-based objection is that *strict* contradiction isn't necessary for a conflict between forces. Suppose, for example, that one force forms the belief that a girl believes a toy is in a box while another believes she believes it's in a basket. As noted above, there isn't a strict contradiction between these beliefs. However, there *is* a conflict between them. A toy can

only be in one place at a time, so the belief that a girl believes a toy in one location is incompatible with the belief that she believes it's in a different location. The two beliefs would also have different behavioral implications. Suppose, for example, that the toy is in the basket, and I want to help the girl find it. If I believe she believes the toy is in the box, I'll steer her in a different direction. If I believe she believes it's in the basket, on the other hand, I'll leave her to her own devices. So, though they don't strictly contradict each other, the two beliefs do conflict with each other in ways that could lead to competition over control of the body.

Arguing that dual processes generate outputs of different mental state types or with different subject matters is not the only way to challenge the claim that the forces in normal human beings come into conflict with each other. Another possible strategy is to deny that human beings are genuinely *committed* to the outputs of both processes in a dual process pair. Think about the last time you made a decision. During your decision-making process, you probably entertained a number of options; you made provisional arguments for various alternatives, 'trying them on for size.' However, you weren't *committed* to all these options. In fact, you were only committed to *one* of them: the one on which you ended up acting. Similarly, it might be argued, human beings aren't committed to the outputs of both dual processes. Rather, they're just committed to the output they actually express. In dual process cases, only one of the outputs is actually expressed, so the human being is only committed to that output. The other output is like the discarded option in the decision-making case; though it might be *entertained* by the human being, he doesn't ever actually *commit* himself to it.

One problem with this kind of objection is that Felix / *dybbuk*-Oscar *also* often expresses only one output. There's a limit to how many behaviors a single body can perform at once and, like the normal human being, Felix / *dybbuk*-Oscar only has one body to work with. As a result, Felix and *dybbuk*-Oscar sometimes have to battle it out for control over the body. When this happens, only one of their (conflicting) outputs is behaviorally manifested. The fact that normal human beings sometimes express only one of a pair of dual process outputs does not, therefore, differentiate them from Felix / *dybbuk*-Oscar. More importantly, the fact that only one of Felix / *dybbuk*-Oscar's behaviors is expressed in certain cases doesn't seem to cast any doubt on the conclusion that Felix / *dybbuk*-Oscar isn't a single, unified person. If it doesn't challenge Felix / *dybbuk*-Oscar's non-personhood, though, why think it challenges the normal human being's?

Another problem with this particular objection is that it's not always *true* that only one of the outputs of a pair of dual processes is expressed. Think, for example, about the IAT studies discussed above. As those studies show, the outputs of implicit and explicit attitude-formation processes can both be behaviorally manifested. Another good example is the mindreading case. There, the outputs of both mindreading processes are expressed by the mindreader. They're expressed in different ways – one is expressed *verbally* while the other is expressed in *non-verbal behavior* (looking time) – but they are both expressed.[21]

A further response to the current objection is that, at best, it only shows that human beings aren't in *synchronic* conflict with themselves. If successful, the objection shows that the forces in human beings aren't in conflict with each other at a given time.

[21] This reading of the empirical findings has also recently been endorsed by Apperly & Butterfill (2009).

However, it doesn't show that they don't conflict *over* time. The processes in parallel dual process pairs compete to produce the response that is consciously expressed. These competitions are sometimes resolved in favor of one of the processes and sometimes in favor of the other. Because the two processes can produce conflicting outputs, this can result in diachronic inconsistency. For example, when conflict is resolved in favor of the conscious moral judgment-generation process, participants tend to say that it's morally acceptable to take action in trolley cases but, when the unconscious process wins out, they say it's unacceptable. Though participants in these scenarios don't endorse conflicting beliefs simultaneously, they *do* endorse conflicting beliefs. In the end, then, there does seem to be genuine conflict between the forces in human beings; it's just manifested *dia*chronically rather than *syn*chronically.

Now, this response might face an objection of its own. People change over time. For example, 7-year-old me had no doubt that I had free will (though I might not have known it by that name) while 30-year-old me isn't so sure. As this illustrates, there are diachronic conflicts that don't – or, at least, don't *clearly* – challenge claims to personhood. For another example of this kind of phenomenon, imagine that the rapture occurred on May 21, 2011, as some had predicted. Now imagine a particular individual – call him Christopher – who was extremely skeptical about the rapture predictions. Indeed, he was skeptical about the existence of God at all, and had written extensively about his skepticism. Before the rapture, Christopher was an ardent atheist. However, even the most ardent of atheists would reconsider when faced with evidence of the rapture. So, when the rapture goes off as predicted, Christopher converts from his pre-rapture ardent atheism to a similarly ardent theism. Many of Christopher's post-rapture commitments

differ from his pre-rapture commitments. To take an obvious example, his pre-rapture belief that God doesn't exist conflicts with his post-rapture belief that God does exist. However, we don't tend to think that post-rapture Christopher is a different *person* from pre-rapture Christopher. Instead, we say that (the same) Christopher has undergone a change.

The conflict in the dual process cases is, however, importantly different from the conflict in cases like this religious conversion case. In the religious conversion case, there's a logical progression from one set of commitments to another. Christopher's commitments change because his *evidence* changes. Though his two commitments might seem inconsistent in *isolation*, they aren't inconsistent in *context*. The overall set of commitments – including both beliefs and the evidence that links them – is coherent. The same is not true of the inconsistency in dual process cases. In dual process cases, individuals don't change their beliefs in response to new evidence. Rather, they change them because they have shifted from use of one process to use of another. A different force wins out than had won out previously. In such cases, there needn't be – and often isn't – a logical progression from an old belief to a new belief; even in context, the beliefs don't cohere with each other. Though changes in commitments like those in the religious conversion case aren't evidence of a genuine conflict within the human being, then, changes like those in dual process cases are.

2.3 The Analogy between Felix / Oscar and Felix / *Dybbuk*-Oscar

So, normal human beings don't seem to differ from Felix / *dybbuk*-Oscar in ways that prevent us from extending our personhood-based conclusions about Felix / *dybbuk*-Oscar to them. Though there are obviously some differences between the two types of beings,

these differences don't break the analogy I've drawn between them. My main conclusion – that human beings aren't single, unified persons – isn't out of the woods yet, though. Recall that there are two steps to the argument for this conclusion:

(1) Felix / *dybbuk*-Oscar is not a single, unified person.

(2) Normal human beings are like Felix / *dybbuk*-Oscar (in personhood-relevant ways).

Though I've defended the second step of this argument, an objector might still challenge the first. According to this kind of challenge, Felix / *dybbuk*-Oscar is a single, unified person so, even if we *are* like Felix / *dybbuk*-Oscar, that doesn't mean we aren't single, unified persons.

Recall that I started the chapter with an introduction to Felix and Oscar as distinct human beings then moved to the discussion of Felix / *dybbuk*-Oscar. As distinct human beings, Felix and Oscar are indisputably not the same person. Therefore, another way to frame the current objection is in terms of a disanalogy between the pairing of Felix and Oscar (Felix / Oscar) and Felix / *dybbuk*-Oscar. According to this framing of the objection, there are differences between Felix / Oscar and Felix / *dybbuk*-Oscar that render them disanalogous. As a result, though Felix / Oscar isn't a single, unified person, Felix / *dybbuk*-Oscar is.

The most salient difference between Felix / Oscar and Felix / *dybbuk*-Oscar is a physical difference. Felix / Oscar occupies two bodies while Felix / *dybbuk*-Oscar occupies only one. Perhaps this physical difference is relevant to personhood. Maybe multiple personalities in multiple bodies are separate persons while multiple personalities in the same body are not. Like the claim that unconscious mental states are different in kind from their conscious counterparts, there are two possible routes to this conclusion:

(1) Single-bodied multiple personalities differ in kind from multiple-bodied multiple personalities *in virtue of their single-bodiedness*.

(2) Single-bodied multiple personalities differ in kind from multiple-bodied multiple personalities in virtue of something *other* than their single-bodiedness.

According to the first claim, it is the fact that the personalities share a body *itself* that makes them a single person; simply by virtue of being housed in the same body, the personalities are merged into a single person. This route to the conclusion seems, to me, to be rather unmotivated. Refer back to the scrambling teletransporter case. In that case, Felix and Oscar stumble into a teletransporter that scrambles their bodies, but leaves them psychologically distinct. The separate personalities they had before they wandered into the machine still exist, but they are now housed in one body instead of two. There's nothing incoherent about this kind of story; we can easily imagine a scenario in which two persons get trapped in one body. This suggests that sharing a body is not enough to make two distinct personalities a single person. The fact that Felix / *dybbuk*-Oscar exists in one body and Felix / Oscar exists in two doesn't, in and of itself, mark a personhood-relevant difference between them.

The obvious fall-back position for the objector is the second of the above two routes to the objection. According to this position, single-bodied multiple personalities differ from multiple-bodied multiple personalities in ways other than their single- and multiple-bodiedness. And, unlike single- and multiple bodiedness, these other differences *do* mark a personhood-relevant difference between them. Here's a possible way to cash this out. As we've already seen, the processes in dual process pairs can come into conflict with each other. It seems as if there has to be something that adjudicates these conflicts. Whether the adjudication happens before either process is initiated or after they have both

produced outputs, it seems as if there has to be some sort of controller that decides between them. This controller seems to link Felix and *dybbuk*-Oscar to each other in a way that Felix and Oscar are not linked. As a result, Felix and *dybbuk*-Oscar also seem to be a single, unified person in a way that Felix and Oscar are not.

This version of the objection is, however, problematic. An initial problem with it is that it's actually *not* obvious that there has to be a third-party controller that adjudicates conflicts between the processes in dual process pairs. To see why not, think about the last disagreement you had with a friend or family member. On some occasions – say, when you've been arguing for a while with no resolution in sight – you might ask someone else to weigh in on the disagreement. Much of the time, though, you can resolve such disagreements without any third-party intervention. Whether by rational persuasion, physical force, or some more creative method, one of the parties to the dispute is convinced by the other (or the two parties strike a compromise), and the disagreement is resolved. Something analogous to this could be going on in the dual process case. Of course, the methods would be different (e.g. higher neural activation levels might stand in for physical force), but it's hardly a foregone conclusion that conflicts between cognitive processes have to be resolved by a third, controller process.

Another problem with the objection is that the presence of a controller doesn't seem sufficient to unify multiple personalities into a single person. Think about some of the occasions – mentioned above – in which two parties to a disagreement appeal to a third-party authority. Does this appeal – or willingness to abide by the third-party authority's ruling – make the disputants a single person? It certainly doesn't seem like it. A better description of what's going on is that two separate persons are appealing to a

third person for help resolving a conflict. So, even if there *is* a controller process that adjudicates disputes between the processes in dual process pairs, that doesn't necessarily mean that the processes are part of the same, unified person. Of course, it does mean that there's a connection between the processes. But, as the case just described establishes, this connection isn't the kind of thing that unifies them into a single person.

The conclusion that internal conflict resolution – whether via a third-party controller or not – doesn't necessarily render a being a single, unified person is confirmed by the fascinating case of Krista and Tatiana Hogan.[22] Krista and Tatiana are craniopagus conjoined twins, which means that they are conjoined twins who are joined at the head. Unlike other conjoined twins, they are also neurally connected to one another. A line of neural tissue – described by their neurosurgeon, Dr. Douglas Cochrane, as a thalamic bridge – connects their thalami. Krista and Tatiana have different personalities: one is allergic to canned corn while the other is not, one likes ketchup while the other despises it, one is lighthearted while the other is more serious, one is more of a bully than the other, etc. (Dominus, 2011). Because of their physical condition, they have to negotiate the conflicting impulses these differences in personalities prompt and, because of their neural connection, they often do this internally.[23] So, Krista and Tatiana Hogan internally resolve conflicts between their impulses. Nonetheless, they seem – to their family and to outside observers like me – to be two separate persons. Given their different preferences, beliefs, etc., it just seems inaccurate to describe them as a single person with two sets of torsos, arms, legs, etc.

---

[22] Thanks to Holly Smith and Tim Campbell for bringing this case to my attention.

[23] Remarkably, as a result of their neural connection, Krista and Tatiana seem also to share sensory perceptions. When one drinks a liquid, the other seems to feel it, an injury to one seems to cause pain to the other, each seems to be able to see what the other can see, etc.

If the physical difference between Felix / Oscar and Felix / *dybbuk*-Oscar doesn't break the analogy between them, maybe something about the nature of the beings in each case does. One potentially relevant difference between the two cases is that both personalities in the Felix / Oscar pairing are conscious while only one of the Felix / *dybbuk*-Oscar personalities is.[24] There are a couple of different ways to motivate the claim that this breaks the analogy between them. According to one of them, (1) the presence of an opposing force only challenges a being's single, unified personhood if the force is itself a person, (2) consciousness is necessary for personhood and, therefore, (3) an unconscious opposing force doesn't challenge a being's single, unified personhood.

This objection echoes one of the objections raised in §2.2. And the same responses offered there also apply here. First, as we'll see in Chapters 3-5, there's reason to doubt that consciousness really is necessary for personhood. The unconscious arguably has the characteristics and can perform the functions we associate with personhood. This at least opens up the possibility of unconscious persons. Second, as the nanorobot example in §2.2 suggests, a force doesn't have to be a person itself to challenge the personhood of a being. Though a nanorobot implanted into a human being's brain isn't itself a fully-fledged person, it does prevent the human being in which it's implanted from qualifying as a single, unified person.

Another way to motivate the claim that Felix / Oscar and Felix / *dybbuk*-Oscar are disanalogous is with a unity of consciousness-based argument. Both forces in the Felix / Oscar pairing are conscious, so the pairing as a whole doesn't have a single, unified

[24] Arguably, the Felix / *dybbuk*-Oscar case is too underdescribed for us to state with certainty that it differs from Felix / Oscar in this way. If we wanted to maintain that Felix and *dybbuk*-Oscar are both conscious, we could just move this objection to §2.2; we could say that this marks a difference between normal human beings and Felix / *dybbuk*-Oscar rather than Felix / *dybbuk*-Oscar and Felix / Oscar.

consciousness. In the Felix / *dybbuk*-Oscar pairing, on the other hand, only one force is conscious. This opens up the possibility that that pairing has a single, unified consciousness. According to the current objection, unity of consciousness amounts to unity of personhood. Therefore, beings that have a single, unified consciousness – like the Felix / *dybbuk*-Oscar pairing – are persons in a way that beings that don't have a single, unified consciousness – like the Felix / Oscar pairing – aren't.

There are a couple of different ways to interpret this objection. On one interpretation, the claim is that Felix / *dybbuk*-Oscar's consciousness is consistent, or coherent, in a way that Felix / Oscar's is not. As we've already seen, though, this isn't quite true. At best, Felix / *dybbuk*-Oscar is only *synchronically* consistent. Because Felix / *dybbuk*-Oscar contains a pair of forces that perform cognitive tasks in different (inconsistent) ways, and each of the forces produces some of the being's conscious thoughts, the being's consciousness is diachronically inconsistent. Whether or not Felix / *dybbuk*-Oscar's consciousness is unified at any given time, it's not unified over time.

Another way to interpret the objection is phenomenologically. On this interpretation, the claim is that the Felix / *dybbuk*-Oscar pairing has a *feeling* of unity that the Felix / Oscar pairing doesn't share. An initial problem with this claim is that it doesn't seem to be true. As I've described Felix / *dybbuk*-Oscar, Felix feels an internal conflict with *dybbuk*-Oscar. He recognizes that he behaves in ways that are out of character (for Felix), and experiences a push-and-pull with the other force inside his body (i.e. *dybbuk*-Oscar). Also, even if the claim *were* true, this objection would echo the objection – first discussed in §2.2 – that *feelings* of unity tap into *actual* unity. Just as with the objection it echoes, there's an obvious response to this objection. Our intuitions,

or feelings, about the way things are are frequently fallible and, in this case in particular, we have positive reason to think that they're on the wrong track.

2.4 A Recap of the Argument and a Clarification

The claim that Felix and Oscar are not the same person is uncontroversial; barring unusual circumstances (or theories), two different personalities in two different bodies clearly aren't a single, unified person. The controversial steps in the current argument are, rather, the two analogies that move us from Felix and Oscar to the normal human being:

(1) The analogy between Felix / Oscar and Felix / *dybbuk*-Oscar and

(2) The analogy between Felix / *dybbuk*-Oscar and the normal human being

To challenge the argument, an objector would have to show that one (or both) of these analogies fails to hold. Now, the beings in each pair obviously differ from each other. However, difference alone isn't sufficient for disanalogy. For a difference between two beings to be a genuine disanalogy in the current context, it has to be a *personhood-relevant* difference. It has to give us grounds for thinking that one being in a pair is a single, unified person while the other is not.

In §2.2 and §2.3, I surveyed some possible disanalogies between Felix / Oscar and Felix / *dybbuk*-Oscar, and between Felix / *dybbuk*-Oscar and normal human beings. In §2.2, I asked whether differences in the natures of Felix / *dybbuk*-Oscar and normal human beings constitute genuine disanalogies. Is the fact that Felix / *dybbuk*-Oscar is not fully human, contains two complete, fully-realized persons, or lacks a feeling of unity a personhood-relevant difference? I also asked whether differences in the nature of the conflicts between forces in the beings constitute disanalogies. Do the processes in normal human beings' dual process pairs fail to generate outputs of the same mental state type or

with the same content? Are there differences in the normal human being's commitment to the outputs? If so, do these kinds of differences mark genuine disanalogies? The answer to all of these questions turned out to be no. In all cases, either the alleged differences between Felix / *dybbuk*-Oscar weren't actually differences or they weren't personhood-relevant. If we doubt that Felix / *dybbuk*-Oscar is a single, unified person, we should have similar doubts about normal human beings.

Given this conclusion, the obvious next question to ask is whether we should doubt that Felix / *dybbuk*-Oscar is a single, unified person. Are Felix / *dybbuk*-Oscar analogous to Felix / Oscar, or is there a personhood-relevant difference between them? In §2.3, I addressed these questions. I started by asking whether the physical difference between the pairings – either in and of itself or because it is accompanied by the presence / absence of a third-party controller – is a personhood-relevant difference. Then, I asked whether a difference in the consciousness of the forces involved marked a genuine disanalogy. Here, again, the answer to the questions was no. It's obviously the case that there is a physical difference between the multiple-bodied Felix / Oscar and the single-bodied Felix / *dybbuk*-Oscar, but this physical difference isn't personhood-relevant. Similarly, though *dybbuk*-Oscar might not be conscious, this difference between him and Oscar doesn't seem to mark a genuine disanalogy between them. As with the normal human being and Felix / *dybbuk*-Oscar, we seem to be able to draw the same personhood conclusions about Felix / *dybbuk*-Oscar as we do about Felix / Oscar. As I've noted, Felix / Oscar are indisputably not the same person. The same seems also to be true of Felix / *dybbuk*-Oscar (and, by extension, normal human beings).

Before concluding this chapter, I should highlight a limitation on the current conclusion that has, so far, remained largely implicit. It might be tempting to take my arguments to show that Felix / *dybbuk*-Oscar or normal human beings contain two (or more) distinct persons. Learning that Felix / *dybbuk*-Oscar is not a single, unified person might lead the casual reader to think that Felix's body contains two distinct persons: Felix and *dybbuk*-Oscar. However, this is too quick. As we saw with the nanorobot example, an opposing force in a human body can interfere with the single, unified personhood of that body even if it is not itself a fully-fledged person. So, the fact that Felix / *dybbuk*-Oscar or the normal human being is not a single, unified person doesn't show that either of them contains two distinct persons; it just shows that neither contains a single, unified person. To establish that Felix and *dybbuk*-Oscar – or the processes in dual process pairs – are two distinct persons, we would have to make an additional argument.

In Chapters 3-5, I try to make this additional argument. I argue that the set of unconscious states and processes has the characteristics, and is capable of performing the kinds of functions, we associate with persons. This opens up the genuine possibility that normal human beings not only don't contain single, unified persons, but might actually contain two distinct persons – one conscious and the other unconscious.

CHAPTER 3: ON ANOTHER CONFUSION ABOUT A FUNCTION OF CONSCIOUSNESS

3.1 The Metaphysical Function of Consciousness

Why do we care about artificial consciousness (i.e. consciousness in man-made machines, like robots and computers)? Why does it matter to us whether fetuses or ants or dogs are conscious? Why does it seem important to know whether permanently comatose patients are conscious of their surroundings? One possibility is that we're interested in consciousness for its own sake. We might study artificial consciousness, for example, just because we want to know whether a future version of IBM's supercomputer, Watson, will be able to introspect its correct *Jeopardy!* answers or a descendent of my Roomba will be able to introspect a desire to clean my carpet. However, this explanation doesn't seem to capture everything that's going on. We care whether a future version of Watson could be conscious not just for its own sake, but also because Watson's consciousness or lack of consciousness seems (at least potentially) to have *metaphysical* implications. Conscious beings seem to be candidates for special metaphysical statuses like personhood in a way that unconscious beings aren't. We care whether Watson is conscious not just because we want to know whether it can introspect beliefs, pains, perceptions, etc., but also because we want to know whether it's a potential person.

I ended Chapter 2 with the following conclusion: human beings aren't single, unified persons. There are multiple forces that compete for control over any given human being, and these forces can't be rationalized into a single, unified person. There's an obvious follow-up question to ask about this conclusion: what exactly are these 'forces'? Think back to the nanorobot case I introduced in §2.2. In that case, a limited-capacity nanorobot is introduced into a human brain. Though this robot competes with the human

person for control over the body they share, it's not a person itself. One possibility is that normal human beings are in a scenario like this one; an unconscious 'nanorobot' (or nanorobots) competes with the conscious person for control over the human being's thoughts and behaviors. According to this possibility, there's only one person in each human being; the other force(s) are non-persons. The other possibility is that more than one of the forces is a person. Not only the conscious but also the unconscious qualifies as a fully-fledged person.

These two possibilities have very different implications. We tend to accord persons more respect than non-persons. I have very different moral obligations to the persons in my life (my family, friends, etc.) than the non-persons (my desk or the chair I'm sitting in). We also expect more of them. My sister can be praiseworthy or blameworthy in a way that my car – no matter how thoroughly I anthropomorphize him – just can't. If the unconscious is a person, it enters into fundamentally different relationships with other beings than if it isn't. Classifying (or failing to classify) the unconscious as a person also has implications for the way we track its identity – and the personal identity of human beings as wholes – across time. Though there's some dispute about exactly what are the requirements for diachronic personal identity, it's generally agreed that they differ from the requirements for other types of diachronic identity (e.g. of artifacts). Also, if human beings contain two persons, there are two diachronic personal identity paths for us to track in each human being rather than just one. Given these implications, it really *matters* which of the two possibilities actually obtains. The unconscious has different rights and responsibilities if it's a person than if it isn't, we

have to use different methods to keep tabs on it if it's a person than if it isn't, etc. So, it matters whether the unconscious could be a person or not.

Now, according to the intuition spelled out at the beginning of the chapter, this question has already been settled. If consciousness is necessary for personhood, unconscious persons are automatically ruled out; the set of unconscious states and processes *can't* be anything more than a 'nanorobot.' Given the significance of our answer to this question, though, we shouldn't arrive at it by intuition alone; it merits a more thorough investigation than that. If unconscious beings really are disqualified from personhood, we should be able to give some reason *why* they are disqualified. There should be something that separates unconscious beings from conscious beings, and justifies according them different metaphysical statuses. There should be something *special* about conscious beings that uniquely qualifies them for – and disqualifies unconscious beings from – personhood.

As I use the term, 'consciousness' amounts to *introspectibility*; conscious states are introspectible states, and conscious processes are processes with at least some introspectible mechanisms. So, the obvious place to start looking for a link between consciousness and personhood is with introspection. Is the ability to introspect a prerequisite for personhood? Is it because Watson and my Roomba can't introspect that they fail to qualify as persons? Could the link between consciousness and personhood really be this simple?

Here's a reason to doubt it: introspection is just a way of 'looking at' mental goings-on, and mere *access* to mental states and processes doesn't seem to have metaphysical implications. To see this general point, consider an analogy to computers.

The relationship between introspection and the mind is like the relationship between computer monitors and computers. Computer monitors give us a window into the operations of computers, but they don't seem to affect the *identity* or *status* of computers. My computer seems to be the same computer whether it's hooked up to its regular monitor, a different monitor, a projector, or even nothing at all. Similarly, the presence or absence of a mental 'monitor' (introspection) doesn't seem to affect the identity or status of the mind it might monitor.

For another way to see the same point, consider the following science fiction scenario. Suppose you're suddenly struck with a very thoroughgoing and highly sophisticated cognitive blindness; you can no longer introspect any of your beliefs, desires, etc. Suppose also that, despite your (first-person) mindblindness, you continue to interact with other people and your surroundings just as you had before. Aside from the ability to introspect, the mindblindness seems to have spared all your cognitive capacities. You perform the same tasks and produce the same behaviors with the same fluency and nuance as before the blindness struck. The only difference is that you can no longer consciously 'see' what your mental states and processes are up to. Now suppose (as is presumably the case) that you were a person before the blindness struck. Do you think you've now lost that status? Are you no longer a person? It doesn't seem like it. If you were a person *before* the blindness struck, you seem still to be one *after* it.

Now, there are a couple of objections that might be raised here. First, the case I just described rests on a pretty big assumption: that it's possible to perform the same functions when mindblind as when not-mindblind. This assumption, it might be objected, is untenable. We have to be able to introspect in order to perform many important

cognitive functions, so mindblind versions of us *wouldn't* be able to do everything non-mindblind versions can do. Second, so far, I've glossed over the issue of sensations. Unlike computers or Roombas or IBM's Watson, human beings can feel pain. In ethics, this capacity to suffer is often considered very important. Indeed, it's sometimes explicitly invoked as a condition of personhood. For example, Cranford & Smith (1987) cite lack of sentience as a reason to deny the permanently comatose personhood. Now, it might be argued that, to really *feel* pain, we have to be able to introspect it. So, even if we concede that introspection of beliefs, desires, etc. isn't necessary for personhood, we might still think introspection of sensations – or, more specifically, introspection of *pain* – is.

Let's consider each of these objections in turn, starting with the second. How plausible is it that the capacity to feel pain is required for personhood? Our intuitions about permanently unconscious patients, like the ones in Cranford & Smith's (1987) argument, aren't a good guide here. Permanent unconsciousness is *radical* unconsciousness. The permanently unconscious don't just lack sentience; they also lack the capacity for self-awareness, decision-making, motor control, etc. So, even if the permanently unconscious *aren't* persons, we can't necessarily trace their lack of personhood to an absence of sentience; there are just too many uncontrolled variables.

Better guides to the importance of sentience are our intuitions about cases of *selective* insentience. As it turns out, there are real-world examples of selectively insentient individuals for us to consider: patients with congenital insensitivity to pain with anhidrosis (CIPA). As the label suggests, CIPA patients are unable to feel pain (or sweat). In all other ways (barring comorbid conditions), however, they are just like non-

patient populations. Does CIPA affect a patient's claim to personhood? It might be hard to form an opinion about this based on just a description of the condition. Take a moment, then, and look up the documentary, *A Life without Pain*. Watching even a couple of minutes of the life of a CIPA patient will – I'd wager – convince you that she has no less a claim to personhood than you or I do. If this is right, though, the capacity to feel pain isn't necessary for personhood.

Though it isn't necessary for personhood, there does seem to be *some* connection between the ability to feel pain and personhood. Before moving on – and to fully dispel the notion that the connection is *necessity* – let's try to spell out what this connection is. One thing that's arguably necessary for personhood is having moral value; part of what it seems to mean to be a person is to be worthy of moral consideration. One way to acquire moral value is to be capable of feeling pain. Because I can feel pain, avoiding injury to me is a moral constraint on your behaviors. Similarly, if we think animals can feel pain, we have more moral compunction about injuring them than if we don't think they can feel pain. So, the capacity to feel pain seems to be a sufficient condition for moral value which, in turn, is (arguably) a necessary condition on personhood. As this suggests, there is *a* connection between the capacity to feel pain and personhood, but it's not direct necessity. The capacity to feel pain is a sufficient condition for a necessary condition on personhood, not itself a necessary condition on personhood.[25]

Having addressed the second objection, let's turn to the first. The idea here is that, though introspection isn't *directly* necessary for personhood, it *is* indirectly necessary.

[25] Importantly, the capacity to feel pain isn't the only way to acquire moral value. Beings can also acquire moral value by having interests, engaging in certain kinds of complex cognition, etc. As we'll see in Chapter 4, unconscious beings can have these kinds of characteristics and capacities. So, the requirement that persons have to have moral value doesn't rule out the possibility of unconscious persons.

Introspection is necessary for performance of some further cognitive function which, in turn, is necessary for personhood. What could this further function be? Before attempting to answer this question, we should distinguish it from another question we might ask about the function of consciousness. Like other features of biological beings, consciousness is thought to have evolved. Assuming it isn't a spandrel (an evolutionary byproduct of an adaptive feature), if consciousness evolved, it performs some adaptive function. When we ask about the function of consciousness, we could be asking about *this* kind of function.

At first glance, questions about this adaptive function of consciousness (call it the 'biological function of consciousness') might seem to have the same answer as questions about the function of consciousness in which we're currently interested (call this the 'metaphysical function of consciousness'). What makes consciousness necessary for personhood might seem to be the same thing that makes it evolutionarily adaptive. It's important to recognize, though, that these are two separate questions, and they could have very different answers. After all, the two types of functions have different satisfaction conditions.

An obvious difference between them is that the metaphysical function of consciousness has to be the kind of thing that could be necessary for personhood while the biological function does not. All that's required for something to count as a biological function of consciousness is that it confers (or conferred) a selective advantage on conscious beings. And functions can confer a selective advantage without being the kinds of things that could be necessary for personhood. Take opposable thumbs and grasping, for example. The capacity of opposable thumbs to grasp things confers a selective

advantage on opposable-thumbed beings. Of course, though, having opposable thumbs isn't necessary for personhood.

Another difference between the biological and metaphysical functions of consciousness is that the metaphysical function has to be unique to consciousness while the biological function does not. Recall from above that the metaphysical function of consciousness is a function for which consciousness is necessary; consciousness is necessary for the metaphysical function which, in turn, is necessary for personhood. Consciousness doesn't have to be necessary for the biological function in the same way. To see this, think again about opposable thumbs and grasping. Opposable thumbs enable us to grasp better than we otherwise could, but they aren't the *only* way to grasp; we can also grasp (if not as dexterously) with our toes, teeth, etc. (Nichols & Grantham, 2000). Features of a being can confer a selective advantage on the being just by enabling it to perform a function better than it would otherwise be able to. Unlike the metaphysical function, the biological function of consciousness need only be *quantitatively* – not necessarily *qualitatively* – different from functions that can be performed unconsciously.

These distinctions between the biological and metaphysical functions of consciousness highlight two criteria a candidate must satisfy to qualify as the metaphysical function of consciousness:

(1) The function must be necessary for personhood.

(2) Consciousness must be necessary for performance of the function.

Now that we've spelled out these criteria, we can start to ask what the metaphysical function of consciousness (hereafter, just 'the function of consciousness') might be.

Which cognitive functions are plausibly necessary for personhood? And is consciousness required for performance of these functions?

I begin, in §3.2, by explaining in more detail how to evaluate candidates for the function of consciousness. Applying the second criterion of the function of consciousness is trickier than it looks. The complication is that there isn't just one way for consciousness to be necessary for performance of a function, so there isn't just one type of necessity claim (i.e. claim that consciousness is necessary for performance of a cognitive function). Different types of necessity claims entail different empirical predictions, so we have to use different evidence to assess them. In §3.2, I tease apart the different kinds of necessity claims, and explain exactly what kind of empirical evidence is relevant to assessing each of them. Next, in §3.3, I identify some candidates for the function of consciousness. I close the chapter, in §3.4, by measuring each of these candidates against the first criterion of the function of consciousness. §3.2, §3.3, and §3.4 lay the groundwork for Chapters 4 and 5. In those chapters, I test whether any of the candidates that survive this chapter satisfy the second criterion of the function of consciousness. Is consciousness necessary for any of them? Relatedly, do any of them qualify as the function of consciousness?

3.2 Types of necessity claims

Suppose that you want to take a calculus class. When you go to enroll, you'll probably have to confirm that you've already taken trigonometry. Having taken trigonometry is necessary for enrolling in calculus. Call this kind of necessity 'prerequisite necessity.' To say that *x* is prerequisitely necessary for *y* is to say that *y* can't obtain without *x* first having obtained. Now suppose that you have successfully enrolled in the calculus class,

and are in the process of learning calculus. Finding the derivative of a function figures heavily in calculus, so part of what it means to learn calculus is to learn how to find the derivative of a function. Knowing how to find the derivative of a function is necessary for knowing calculus. The kind of necessity involved here is, however, different from the kind of necessity in the enrolling-in-a-calculus-class case. Knowing how to find the derivative of a function isn't a *prerequisite* for knowing calculus; rather, it's *part* of knowing calculus. Call this second type of necessity 'constitutive necessity.' To say that *x* is constitutively necessary for *y* is to say that part of what is required for *y* to obtain is for *x* to obtain.

In the case of cognitive processes, constitutive necessity can be further broken down into two subtypes. As mentioned in the Introduction, there are three main parts to cognitive processes:

(1) Inputs

(2) Outputs

(3) Mechanisms that translate inputs to outputs

One way for consciousness to be constitutively necessary for performance of a function is for the *inputs* to the process that performs the function to be necessarily conscious. Call this type of constitutive necessity 'input constitutive necessity.' Another way for consciousness to be constitutively necessary for performance of a function is for one (or more) of the process' *mechanisms* to be necessarily conscious. Call this kind of constitutive necessity 'mechanism constitutive necessity.'

So, there are three main types of necessity claims:

(1) Prerequisite necessity claims

(2) Input constitutive necessity claims

(3) Mechanism constitutive necessity claims

According to the first type of necessity claim, to perform a particular function, a being has to have been conscious of something related to the function in the past. An example of this kind of necessity claim is Humphrey's (1983) claim that, in order to draw connections between mental states and behaviors in others, we first have to have introspected similar connections between mental states and behaviors in ourselves.[26] According to the second type of necessity claim, the inputs to the process that performs the function have to be conscious. McGinn's (1982) claim that we can only monitor mental states of which we are introspectively aware is an example of this kind of necessity claim. Finally, according to the third type of necessity claim, one (or more) of the mechanisms of the process that performs the function has to be conscious. An example of this kind of necessity claim is the neo-Kantian idea – described by Suhler & Churchland (2009) – that the mechanisms of deliberative processes have to be conscious.[27]

The different types of necessity claims make different empirical predictions. For example, Humphrey's (1983) prerequisite necessity claim predicts that we won't be able to explain behaviors in mental state terms if we haven't first introspected similar connections between mental states and behaviors in ourselves. McGinn's (1982) input necessity claim, on the other hand, predicts that we won't be able to monitor introspectively inaccessible mental states. As a result, different types of evidence are relevant to assessing each type of necessity claim. In a case like Humphrey's – where the

[26] As we'll see in Chapter 4, this is just one way to interpret Humphrey's (1983) claim. It can also be interpreted as an input constitutive necessity claim.

[27] For more detailed discussions of the necessity claims referenced here, see Chapter 4.

claim is a prerequisite necessity claim – we should be on the lookout for cases in which performance of the function isn't preceded by the related type of conscious processing. In the case of an input constitutive necessity claim like McGinn's, by contrast, we should look for processes that perform the function on unconscious inputs.

Given that different types of evidence are relevant to different types of necessity claims, it's important to be clear about exactly what type of claim we're assessing in any given case. There are a couple of reasons for this. The first is theoretical. We want to know not just *that* a particular piece of evidence confirms or disconfirms a claim, but also *why* it does so. To see why the evidence disconfirms (or confirms) the claim, we have to know exactly what the claim is supposed to be. We have to know why consciousness is supposed to be necessary for performance of the function.

A second, and perhaps more important, reason is a practical one. There are practical constraints on empirical work (e.g. it's simpler to design experiments with supraliminal than subliminal prompts), and these constraints limit the evidence we have available to us (e.g. there are many more studies with supraliminal prompts than subliminal ones). If we aren't precise about where consciousness is supposed to enter a process, the only evidence we can use to test necessity claims is evidence that involves thoroughly unconscious beings. If we *are* precise, on the other hand, we can also call on other sources of evidence. For example, when assessing an input necessity claim, we can use any and all evidence of processes with unconscious inputs (even if those processes have conscious mechanisms or are preceded by conscious processing). Given the practical constraints on empirical research into unconscious states and processes, it's important to not needlessly restrict the class of potential sources of evidence.

3.3 Candidates for the Function of Consciousness

Imagine a pair of businessmen. Call them 'Bateman' and 'Allen.' Bateman and Allen have a not-so-friendly rivalry with each other. At the moment, Allen seems to be coming out ahead in this contest: his numbers are up at work, he's fitted out in the finest suits, and his business cards are just the right shade of off-white. This has prompted a festering resentment in Bateman, who usually – though not always – manages to keep it simmering below the surface. Consider two scenarios. In the first, Bateman notices that Allen is about to walk past his desk. Seized by a sudden childish vindictiveness, he decides to trip him. As Allen walks by, Bateman sticks out his foot. In the second scenario, Bateman is engrossed in his work when Allen is about to pass his desk, and doesn't see him coming. By coincidence, he has a leg spasm just as Allen is walking by. The spasm shoots his foot out into Allen's path.

Now generalize the second case. Imagine a being – call him 'RoboBateman' – whose actions always come about in this automatic kind of way, whose behavioral repertoire is a collection of spasms and reflexes. Does RoboBateman strike you as the kind of thing that can qualify as a person? Probably not. At least part of what distinguishes persons from non-persons, it seems, is that at least some of their thoughts and behaviors are controlled. Why is control relevant to personhood? The most obvious connection is via responsibility. Control plays a central role in responsibility – a being is *responsible* for a thought or behavior provided that it exerts a particular kind of *control* over that thought or behavior – and responsibility seems to be required for personhood. Part of what distinguishes persons like me from non-persons like the tree outside my window or the chair I'm sitting in or the laptop on which I'm writing this chapter is that

persons can be responsible for the things they think, say, and do. So, we have our first candidate for the function of consciousness: responsibility.

Another candidate is rationality. Cast your mind back to the first Aristotle class you took as an undergraduate. Among the things you probably learned in that class was that Aristotle characterized man as a 'rational animal.' While some of Aristotle's ideas – like the notion that the brain is just an elaborate cooling system for the body – have long since been rejected, the idea that rationality and personhood are linked has endured. It crops up in Kant (1788/1997), whose justification for the categorical imperative depends on the assumption that persons are rational beings. And it recurs in many contemporary accounts of persons, like the one Warren (1973) employs in her discussion of the moral dimension of abortion. Indeed, the rationality-personhood link is so popular that Dennett (1976) cites it as one of six common themes of personhood. In a handy survey of the personhood literature, Dennett identifies six oft-invoked criteria of personhood. One of these six criteria is rationality.

Dennett's (1976) survey also supplies us with some other candidates for the function of consciousness. Among Dennett's (1971) best-known contributions to philosophy is the concept of the intentional stance. To take the intentional stance toward something is to treat it as if it's an intentional system. An intentional system, in turn, is a system whose behavior can be understood in terms of intentional states like beliefs and desires. Suppose, for example, that you witness a man pull out a gun, point it at a passing pedestrian, and say, 'Your money or your life.' You can explain this behavior in terms of a belief and a desire: the man points the gun at the pedestrian and says what he says because he *wants* money from the pedestrian, and *believes* that waving around firearms

and threatening clichés is a way to get it. If you explain the would-be mugger's behavior in this way, you're treating him as an intentional system; you're taking the intentional stance toward him. To qualify as a person, a being has (at least sometimes) to behave in ways that are explainable in these kinds of terms. So, one criterion of personhood is the capacity to be an *object* of the intentional stance.

Another, closely related, criterion of personhood is the capacity to be a *subject* of the intentional stance. When I encounter another human being or a dog or the moving shapes in Heider & Simmel's (1944) famous video, my natural response is to explain their behaviors in terms of intentional states. Heider & Simmel's small triangle is moving in the direction it is, at the speed it is, because it *wants* to escape the large triangle, and *believes* that moving in that direction at that speed will enable it to make its getaway. My sister's dog is barking at the door because she *wants* to go outside, and *believes* that barking at the door will prompt my sister to let her out. This ability to explain behaviors in terms of intentional states is another possible distinguishing feature of persons. Persons seem not only to be the kinds of things toward which the intentional stance can be *taken*, but also the kinds of things that can *take* the intentional stance.

Other candidates for the function of consciousness that crop up in Dennett's (1976) discussion are language and self-monitoring.[28] First, as Dennett points out, many personhood theorists want to exclude non-human animals from the class of potential persons. A quick and easy way to do this is to stipulate that persons have to be capable of language. Because few animals seem to have the capacity for language, counting language as a criterion of personhood seems to rule out almost all non-human animal

[28] The final commonly-cited criterion of personhood on Dennett's (1976) list is consciousness itself. I don't address this criterion here because I've already shown – in §3.1 – that consciousness is not genuinely necessary for personhood.

persons. Second, persons seem to be able to monitor their cognition in a way that non-persons can't. They seem capable of reflecting on – and endorsing or rejecting – their motivations, and of intervening in their mental processes to correct or terminate them. This raises the possibility that part of what it means to be a person is to have higher-order oversight over mental goings-on.

All the functions I've identified so far are what we might describe as 'local functions.' These functions only have to involve a single type of process; at least in principle, the processes that perform them could operate in isolation of other processes. There's also another general category of candidates for the function of consciousness. Call the functions in this category 'global functions.' A defining characteristic of persons is that they are psychologically *unified*, or *centralized*. There seems to be some sort of centralized authority in persons that enables them to coordinate their cognitive processes. This centralized authority facilitates information-sharing between the processes, and enables them to work together in pursuit of big-picture objectives. These broadcasting and coordination functions are what I'm describing as global functions. In contrast to local functions – which involve individual process types – these functions involve the system as a whole (or some sizeable subset of the system).

3.4 Necessary Conditions on Personhood

Dennett (1976) and Warren (1973) both advance lists of possible criteria of personhood. Both also advance the same disclaimer about their lists: though the functions on the list are all *possible* criteria of personhood, they might not all *really* be necessary for it. The fact that Dennett and Warren both (independently) feel compelled to include this disclaimer highlights just how difficult it is to establish that any given capacity is

genuinely necessary for personhood. There are lots of things that *could* be necessary for personhood, but it turns out to be pretty difficult to establish that anything *definitely* is.

Fortunately, I think we can sidestep most of this controversy. What we're trying to figure out is whether unconscious beings are automatically excluded from the class of persons. For this purpose, it would be handy to have a complete list of established criteria of personhood, but it isn't strictly necessary. Recall that there are two criteria a function has to satisfy to be the function of consciousness:

(1) The function has to be necessary for personhood.

(2) Consciousness has to be necessary for performance of the function.

Failing to satisfy *either* of these criteria is sufficient to take a function out of the running to be the function of consciousness. So, here's a strategy that's available to us: determine which of the candidates for the function of consciousness is *plausibly* necessary for personhood then check whether consciousness is necessary for any of these candidates. If consciousness isn't necessary for any of the plausible candidates, we have an answer to our main question – are unconscious beings automatically excluded from the class of persons? – without ever having to determine which functions are definitely necessary for personhood.[29]

For the most part, the functions identified in §3.3 do seem plausibly to be necessary for personhood.[30] For example, persons do seem to have to be capable of bearing responsibility for some of their thoughts and behaviors, and they do have to have

[29] Of course, there's always the possibility that consciousness *is* necessary for performance of one of the plausible candidate functions. However, I think we can cross that particular bridge if we come to it.

[30] My defense of this claim is exceptionally cursory. The reason for this is that my endorsement of the claim is a concession to my opponent. My ultimate goal is to show that none of the proposed candidates for the function of consciousness actually qualifies as the function of consciousness. So, it would actually *help* my case if the claim I'm defending here were false.

the capacity to coordinate multiple cognitive processes in the service of big-picture projects. However, there's one possible exception to this rule: language. The justification Dennett (1976) cites for counting language as a criterion of personhood is that doing so handily excludes most non-human animals from the class of persons. But this justification hardly seems principled. Suppose I decide – antecedent to any argumentation – that I want to exclude blue-eyed beings from the class of persons. I can jerry-rig my definition of personhood to exclude the blue-eyed. Indeed, I can even do so sneakily, by citing as my explicit criterion something that's unique to the blue-eyed – say, some perfectly-correlated genetic feature – rather than blue-eyedness itself. This doesn't seem, however, to be a good way of arriving at a criterion of personhood. Separating persons from non-persons then *post hoc* justifying the boundaries we've drawn isn't the right way to go about things. Instead, we should come up with a principled reason to draw the boundaries where we do. To count language as a necessary condition of personhood, we would have to come up with a better justification than that doing so enables us to exclude certain beings from the class of persons.

Does a better justification exist? That depends on what we mean by 'language.' There are a couple of different ways to understand 'language' in this context. According to the first, narrower understanding, what we mean by 'language' is just verbal expression [or verbal-like expression (e.g. sign language)]. If we have this narrower interpretation in mind, asking whether a being has the capacity for language just amounts to asking whether it's capable of expressing itself verbally (or in sign). According to the second, broader understanding, 'language' is something more general. It's a capacity for thinking in language-like ways, or an ability to engage in sequential, combinatorial

thought. A being that has the capacity for language in this second, broader sense is a being that can think in sentences.

Whether we're justified in counting language as necessary for personhood depends on which of these interpretations we choose. On the first interpretation, language doesn't seem necessary for personhood. The relationship between language in this sense and the mind is like the relationship between *introspection* and the mind. Like introspection, language (in this narrower sense) is just a way of accessing mental states and processes. As we saw in the earlier discussion of introspection, though, mere *access* to mental states and processes doesn't have metaphysical implications. To see this in the current context, think about the case of Helen Keller. When she was nineteen months old, Keller contracted an illness that left her unable to see or hear. Until the age of six, she was almost entirely incapable of communicating. Only then, with the help of the remarkably patient and dedicated Annie Sullivan, did she learn to express herself. Before working with Sullivan, Keller wasn't able to use language and, after working with her, she was. This was indisputably an important transformation. However, it wasn't a *metaphysical* transformation. The sessions with Sullivan didn't transform Keller from a non-person to a person; they just enabled her to *express herself* as a person. If this is right, language in the first sense is not necessary for personhood.

Language in the second sense, on the other hand, might be necessary for personhood. More specifically, language in the second sense might be necessary for performance of some further functions which, in turn, might be necessary for personhood. What kinds of further functions could these be? Baumeister & Masicampo (2010) suggest one possibility. Sequential, combinatorial thought enables us to construct

narratives; because we can combine thoughts with each other and think in sequences, we can string events together into causally-connected chains. Once we have learned to string *actual* events into such chains, it's a small step to stringing *possible* events into causal chains. If we can string possible events into causal chains, we can simulate events that aren't actually happening; we can project ourselves into the past and future, others' perspectives, hypothetical scenarios, etc. As we'll see shortly, this kind of simulationist cognition plays a central role in some of the functions that are plausible candidates for the function of consciousness – namely, rationality and being a subject of the intentional stance.

Given the above conclusions about language, we're left with eight plausible candidates for the function of consciousness – six local candidates and two global candidates:

Local candidates:

(1) Being an object of the intentional stance

(2) Being a subject of the intentional stance

(3) Responsibility

(4) Rationality

(5) Language-like thought

(6) Self-monitoring

Global candidates:

(7) Broadcasting

(8) Coordination

These eight candidates plausibly satisfy the first criterion of the function of consciousness. The next step is to determine whether any of them also satisfies the second criterion. Is consciousness necessary for performance of any of these functions? I answer this question for the local candidates in Chapter 4 then turn to the global functions in Chapter 5.

CHAPTER 4: LOCAL CANDIDATES FOR THE FUNCTION OF CONSCIOUSNESS

4.1 Introduction

In Chapter 3, I settled on six plausible local candidates for the function of consciousness. One of these contenders (language-like thought) is notable primarily for its role in facilitating two of the others (rationality and being a subject of the intentional stance). Language-like thought seems to be necessary for simulation. More specifically, it's *constitutively* necessary; stringing together hypothetical chains of events is *part* of the simulation process. And simulation is what we often use to rationalize and to attribute mental states to others.[31]

This relationship between language-like thought and rationality / mindreading has an important implication: evidence of unconscious rationality / mindreading does double duty as evidence of unconscious language-like thought. We can't simulate without thinking in a language-like way so, when we engage in unconscious simulationist rationalization or mindreading, we're also engaging in unconscious language-like thought. This means that we can determine whether language-like thought is necessarily conscious *en route* to determining whether being a rational being or a subject of the intentional stance is; if rationality and mindreading aren't necessarily conscious, neither is language-like thought. Given this, I fold my discussion of language-like thought into my discussions of rationality and being a subject of the intentional stance.

I take a similar approach with self-monitoring and responsibility. Responsibility ultimately boils down to control; a being is *responsible* for a thought or behavior provided that he exerts a particular type of *control* over it.[32] As we'll see shortly, some of

[31] For defenses of each part of this claim, see §5.2 and §5.4, respectively.
[32] For a defense of this claim, see §5.3.

the candidates for this 'particular type of control' involve self-monitoring; self-monitoring is constitutively necessary for them. This means that, just as we can test whether language-like thought is necessarily conscious by checking rationality and mindreading, we can test whether self-monitoring is necessarily conscious by checking these types of control. As with language-like thought and rationality / mindreading, we can determine whether self-monitoring is necessarily conscious *en route* to determining whether these types of control are.

So, my discussion of local candidates for the function of consciousness can be grouped into three sections. In the first, §4.2, I ask whether either of the intentional stance-related candidates for the function of consciousness – being an object of the intentional stance and being a subject of the intentional stance – satisfies the second criterion for the function of consciousness. In the course of answering this question for the second intentional stance-related criterion (being a subject of the intentional stance), I also answer it for language-like thought. Next, in §4.3, I ask whether responsibility and control – including the kinds of control that involve self-monitoring – are necessarily conscious. Then, in §4.4, I ask whether rationality – and, by extension, language-like thought – can be unconscious. I end the chapter, in §4.5, with a brief summary of the findings of the preceding sections.

4.2 Objects and Subjects of the Intentional Stance

To be an object of the intentional stance is just to be interpretable as an intentional system. You don't actually have to *be* an intentional system; you just have to be *interpretable* as one. This is not a demanding standard. My one-year-old niece, Elise, who can't seem to stay away from the dog's water bowl, is interpretable as an intentional

system. So is the dog, Molly, whose bowl she so regularly tips over. Even the fridge that keeps Molly's water cool can be interpreted as an intentional system. There's a thermostat in the fridge that causes the cooling element to click on when the temperature reaches one threshold and off when it hits another. This behavior can be explained in terms of a *desire* to keep the inside of the fridge within a certain temperature range, and a *belief* that clicking the cooling element on or off will satisfy this desire.

Given this understanding of what it means to be an object of the intentional stance, it seems clear that the capacity to be an object of the intentional stance isn't limited to conscious beings. Indeed, we've already encountered some unconscious objects of the intentional stance. Depending on your position on non-human animal consciousness, you might think Molly is an unconscious object of the intentional stance. Unless you have some very strange views indeed, you'll *definitely* think the fridge is an unconscious object of the intentional stance. Anything that can fruitfully be treated as having intentional states is something toward which we can adopt the intentional stance. And there are lots of unconscious things we can fruitfully treat as having intentional states.

In response to this, it might be argued that Dennett's (1976) understanding of 'intentional systems' is simply too weak. Rather than thinking of intentional systems as systems that are merely *interpretable* as having intentional states, we should think of them as systems that actually *have* intentional states. They aren't just beings to which we can attribute propositional attitudes, but beings of whom such attributions are true.

Even on this more demanding interpretation, though, unconscious beings don't seem to be disqualified from being objects of the intentional stance. To see this, refer

back to the discussion of Searle (1992) in Chapter 2. As we saw there, *contra* Searle, there doesn't seem to be a necessary link between consciousness and intentionality; states can be intentional even if they are unconscious. Even if we stipulate that objects of the intentional stance are systems that actually have – rather than are just interpretable as having – intentional states, then, unconsciousness doesn't disqualify beings from being objects of the intentional stance. Under either the weaker or the stronger interpretation, being an object of the intentional stance isn't necessarily conscious, and it doesn't satisfy the second criterion of the function of consciousness. It can't be the function of consciousness we're looking for.

What about being a *subject* of the intentional stance? To be a subject of the intentional stance is to treat others as bearers of intentional states. Another way to describe this activity is as attributing intentional states to others, or mindreading. So, to be a subject of the intentional stance, a being has to be able to mindread. Many theorists take mindreading to be necessarily conscious. For example, in their respective investigations of the function of consciousness, Humphrey (1983) and Baumeister & Masicampo (2010) both describe mindreading as uniquely conscious. If this is true, only conscious beings can be subjects of the intentional stance. So, *is* it true? Is mindreading really unique to conscious beings?

Let's take a look at the arguments that it is, starting with Humphrey's (1983). What distinguishes human beings from spiders or rattlesnakes or rhinoceroses? One difference, Humphrey suggests, is consciousness. When I'm hungry, I'm consciously aware that I'm hungry. I feel my hunger, and can use this feeling to explain why I behave the way I do (e.g. walking to the kitchen and opening the fridge). The rhino, on the other

hand, has no such resources. Though he engages in food-seeking behavior, he can't look inside his mind and plumb the reasons for this behavior.

Another difference between me and the rhino, according to Humphrey (1983), is that I can engage in a particular kind of social psychology. Unlike the rhinoceros, I can attribute mental states to others, explain their behaviors in terms of their mental states, etc. Prompted by this observation, Humphrey constructs a hypothesis about the role of consciousness in social psychology. Without an understanding of mental states, beings needn't be at a *complete* social loss; even in the absence of a window into minds (either their own or others), they can provide behaviorist explanations of behaviors. As the history of behaviorism shows, however, this behaviorist kind of social psychology isn't adequate for all social psychological purposes. To take just one example, the same physical behaviors take on different meanings when they are conducted in pursuit of different goals. Running-after-someone-to-steal-his-wallet and running-after-someone-to-retrieve-a-wallet-he-stole might look identical. To differentiate between them, we have to identify the (different) goals that motivate them. Consciousness, Humphrey hypothesizes, is what enables us to do this. Introspection of our *own* mental states and the ways in which they lead to *our* behaviors provides us with a model for attributing mental states to *others* and explaining *their* behaviors in terms of mental states. So, according to Humphrey, the function of consciousness is something we might describe as mentalistic social psychology. In contrast to behaviorist social psychology – which explains behaviors in behaviorist terms – mentalistic social psychology explains behaviors in terms of mental states. It's the kind of social psychology that requires mindreading.

Humphrey's (1983) hypothesis is an empirical one and, like other empirical hypotheses, it can be tested against the predictions it makes. His claim is that we are able to engage in mindreading because we can introspect the mental state-based explanations of our own behaviors, and extrapolate those explanations to the behaviors of others. This claim suggests the following prediction: mindreading won't occur in the absence of introspection of inner explanations. Now, to introspect inner explanations is not just to introspect inner *states*. Rather, it's to introspect inner states *as* explanations of behaviors. This involves not just conscious *detection* of mental states, but also conscious *reasoning* about them. To introspect inner explanations is to introspect not only states, but also the roles they play in producing behaviors.

How can we test Humphrey's (1983) claim against this prediction? One way to do it is by looking for cases in which beings are capable of mindreading but not conscious reasoning about their own mental states. Do such cases exist? Finding them is a tricky proposition. Engaging in third-person mindreading and consciously reasoning about your own mental states are both sophisticated tasks, and paradigmatic practitioners of one – like normal adult human beings – tend also to be paradigmatic practitioners of the other. To find a being that is capable of mindreading but not conscious reasoning, we have to look to the margins of social psychology.

One example of a social psychologist at the margins is the schizophrenia patient with passivity symptoms. In some cases, patients with schizophrenia feel a lack of ownership or control over their thoughts or behaviors (passivity symptoms). For example, a patient with schizophrenia might feel as if his movements are being controlled by an external 'puppetmaster,' or his mind is devoid of thoughts. Hurlburt reports that one such

patient, Joe, "could not describe *any* aspects of his inner experience" (1990, p. 208, emphasis in original). Though they lack access to their own states, however, these patients are not impaired at mindreading. A recent meta-analysis of mindreading in schizophrenia confirmed that patients with passivity symptoms perform at normal levels on mindreading tasks (Sprong et al., 2007). This finding is robust across both patients in remission and symptomatic patients (Frith & Corcoran, 1996). Combined, these findings suggest that symptomatic patients with schizophrenia with passivity symptoms can engage in mindreading, but can't introspect their own mental states – let alone consciously reason about them (Nichols & Stich, 2003). These patients seem, therefore, to be counterexamples to Humphrey's (1983) hypothesis.[33]

Before we consider the case closed, though, we should note that Humphrey (1983) has a response to this available to him. He could say that my objection treats his claim as a *constitutive* necessity claim when it should be treated as a *prerequisite* necessity claim. In formulating my objection, I assumed that Humphrey's claim was that conscious reasoning about inner explanations is *part* of the mindreading process. However, there's another possible interpretation. According to this alternative interpretation, when we consciously reason about the connections between our mental states and behaviors, we build up a store of connections. These connections are then available to be deployed – consciously or unconsciously – whenever we need them. On this interpretation, there's a ready explanation of the schizophrenia findings reported above: schizophrenia patients consciously store up inner explanations when in remission

[33] Further evidence for mindreading in the absence of current introspection of inner explanations comes from work with visual extinction patients. For a more detailed discussion of this work, see the below discussion of Baumeister & Masicampo (2010).

then unconsciously deploy them when symptomatic. Consciousness enters the picture – and can exit it – well before any particular mindreading process takes place.

On this picture, mindreading is an automatizable process – rather like driving. When you're first learning to drive, you have to employ a great deal of information consciously. You have to think consciously about where to look, when to shift, etc. Over time and with practice, however, driving becomes automatic for you. You no longer have consciously to think about where to look or when to shift. Similarly, when you're first learning to mindread, you have consciously to inspect the relationships between your own mental states and behaviors, and figure out how to extrapolate them to *others'* mental states and behaviors. Over time and with practice, however, mindreading becomes automatic. You no longer have consciously to model your attributions of mental states to others on the connections you observe in yourself.

Now, the fact that you can (with time and practice) drive a car without thinking consciously about where to look or when to shift doesn't show that introspection of this looking and shifting information isn't necessary for driving; it just shows that it isn't *currently* necessary. Similarly, evidence that schizophrenia patients can mindread without introspecting inner explanation information doesn't show that introspection of that information isn't necessary for mindreading; it just shows that it isn't *currently* necessary. To put this slightly differently, though introspection of inner explanations might not be *constitutively* necessary for mindreading, it is *prerequisite* necessary.

Of course, for this response to be convincing, Humphrey (1983) would have to provide some evidence to support it. What reason is there to believe that we store up inner explanations in this way? Also, even if there was evidence for the automatization

explanation of the schizophrenia case, there are other cases for which this kind of explanation doesn't work. Why not? Because some unconscious mindreading processes draw on inner explanation information that isn't *ever* introspectively accessible. Though the beings that engage in these processes might have introspective access to some inner explanation information, they don't have access to the inner explanations that are operational in these particular processes. If they don't ever have access to these inner explanations, the processes can't be automatized versions of previously conscious processes. If I can't ever introspect a link between a particular set of behaviors and mental states, my unconscious deployment of that link can't be an automatized version of a previously conscious process.

One example of an unconscious mindreading process that draws on non-introspectively accessible inner explanation information is lie detection on the basis of bodily leakage. Bodily leakage occurs when a liar's true beliefs are revealed – always unintentionally and usually without the liar's knowledge – by his nonverbal behaviors. Facial and bodily leakage are both common during deception. While people tend to be aware of (and try to correct for) facial leakage, they are typically unaware of the potential for bodily leakage; they don't realize that there's a connection between their attempts to deceive and certain patterns of bodily movements (e.g. decreased head motion). When trying to *detect* deceit, however, perceivers employ bodily leakage information. Though they aren't able to say why, perceivers interpret individuals who display decreased head movement and other types of bodily leakage as insincere (Choi et al., 2005). In this kind of case, mindreaders don't ever have introspective access to the link between bodily

leakage and deceit. Nonetheless, they are able to use information about this link in their mindreading.[34]

So, Humphrey's (1983) argument for the necessity of consciousness to mindreading doesn't seem to succeed. What about Baumeister & Masicampo's (2010) argument? In their article, Baumeister & Masicampo list some things consciousness *doesn't* seem to do for us. Citing findings that our conscious interpretations and explanations are often wrong (Gazzaniga, 2003; Nisbett & Wilson, 1977; Libet, 1985; Wegner, 2002), they conclude that a commonsense picture of the function of consciousness – according to which it is what enables us to perceive the world and direct our behaviors in it – is probably wrong. One moral we could draw from this is that consciousness is epiphenomenal; if consciousness doesn't do either of the things we commonly take it to do, maybe it doesn't do anything at all. Baumeister & Masicampo take a different tack. If consciousness isn't how we get information from, or exert influence over, the physical world, they suggest, maybe it isn't built for dealing with the *physical* world at all. Perhaps, instead, it's built for dealing with the *social* world.

Among the social functions Baumeister & Masicampo (2010) trace to consciousness is mindreading. Employing the logic described in §4.1, they suggest that consciousness is what enables us to string together hypothetical events into narratives. This capacity enables us to simulate other's perspectives, or 'put ourselves in others' shoes.' As Goldman (2006) compellingly argues – and Baumeister & Masicampo concur – much of our mindreading involves such simulation. The primary way in which we attribute mental states to others is by simulating the processes by which they arrived at

[34] Further confirmation of mindreading in the absence of introspective access to inner explanations might come from work with dogs. Call et al. (2003) have shown that dogs are remarkably good mindreaders, but work by Bräuer et al. (2004) suggests that they aren't very adept at introspection.

those states. If mindreading involves simulation and simulation requires consciousness, Baumeister & Masicampo reason, consciousness seems necessary for mindreading.

Are they right about this? Does simulationist mindreading have to be conscious? The answer here is a fairly resounding no. There's a good deal of evidence that we can mindread unconsciously. We've already encountered some of this evidence. In the above discussion of Humphrey's (1983) argument, we saw that schizophrenia patients with passivity symptoms and normal adult human lie detectors can mindread unconsciously and, in Chapter 1, we saw that young children can do the same. These studies show that the mechanisms of mindreading processes can be unconscious.

Another study – so far, relegated to a footnote – shows that mindreading processes can also take unconscious inputs. Human beings transmit a great deal of information through their facial expressions. The primary method we use to decode – or *attribute* – this information is something called face-based emotion recognition (FaBER). Work with both normal populations and visual extinction patients shows that FaBER can occur entirely unconsciously. First, normal participants display the same behavioral responses (e.g. mimicry) and neural responses (e.g. amygdala activation) to unconsciously-perceived facial emotional stimuli as to consciously-perceived facial emotional stimuli (Choi et al., 2005). In other words, even when they aren't consciously aware that a face is present, normal participants display behavioral and neural evidence of recognition of the emotional information conveyed by the face.

Second, visual extinction is permanent introspective blindness in parts of the visual field. Patients with this condition lack any conscious access to stimuli in the extinct parts of their visual fields so, if they attribute mental states to these stimuli, they

can't be doing so consciously. And they *do* attribute mental states to such stimuli. Like non-patient populations, patients with visual extinction engage in FaBER. More surprisingly, they employ it with emotional stimuli that are presented to their extinct fields. For example, when a happy face is presented to an extinction patient's blind field then another happy face is presented to her non-blind field, she displays priming effects; she is faster at recognizing the emotion in the second face (De Gelder et al., 2005; Williams & Mattingley, 2003).

Now, this kind of effect would occur only if the patient was processing the stimuli's *emotional content*. Why? Presentation of a stimulus only facilitates processing of another stimulus if the two stimuli are congruent.[35] So, if the extinction patient displays priming effects, she must be processing the stimulus that's presented to her blind field *as* emotionally congruent with the stimulus presented to the non-blind field.[36] Of course, if she's processing the stimulus that's presented to her blind field as emotionally congruent to the other stimulus, she must be processing the *emotional content* of the unconscious stimulus. Processing the emotional content of a stimulus is, of course, recognizing the emotions expressed by the stimulus. Therefore, the participants in these experiments were engaging in genuine *unconscious* FaBER (mindreading) processes. This suggests that mindreading can occur in the absence of consciousness not only of mindreading mechanisms, but also the inputs to these mechanisms. *Pace* Baumeister &

[35] As we'll see in an upcoming section, there is one exception to this (reverse priming). This exception isn't relevant for present purposes.

[36] In response to this, it might be pointed out that there's more than one way for two stimuli to be congruent with each other. How do we know that the priming effect in this study was due to *emotional* congruence between the stimuli rather than, say, *physical* or *structural* congruence? Fortunately, the experimenters controlled for this potential confound. Williams & Mattingley (2003) used different images for primes than targets, and the prime images they used were smaller than the targets and presented peripherally. Therefore, the participant must have been responding to similarities in emotional content rather than mere physical or structural similarities.

Masicampo (2010), consciousness doesn't seem to be required at any point in simulationist mindreading processes.[37]

4.3 Responsibility

To determine whether responsibility is necessarily conscious, we first have to get clearer on what it means to be responsible for a thought or behavior. As noted in Chapter 3, responsibility seems to consist in control; I'm responsible for winning a big hand in poker or cheating at a philosophy department mini-golf game provided that I have a particular kind of control over my poker-playing or mini-golf-cheating behaviors. What exactly is this particular kind of control? What kind of control do I have to exert over a thought or behavior to qualify as responsible for it? For an answer to this question, let's revisit the Bateman and RoboBateman examples introduced in Chapter 3. Bateman seems to have control over – and be responsible for – his tripping behavior in a way that RoboBateman is not. What's the difference between the two tripping behaviors? What kind of control does Bateman exert over his behavior, and RoboBateman fail to exert?

A tempting answer is that Bateman could have acted in a way other than the way he did act while RoboBateman couldn't; Bateman seems to have selected his behavior from a range of equally available options while RoboBateman was forced into the behaviors he performed. Historically, however, this way of thinking about control – which Fischer (1987) refers to as 'regulative control' – has run into difficulties. Empirical evidence suggests that the world is deterministic in the following sense: given the initial

[37] The conclusion that mindreading can occur unconsciously has an interesting further implication. Some animal cognition theorists use evidence of mindreading to argue for consciousness in non-human animals (see, for example, Butler & Cotterill, 2006; Edelman & Seth, 2009). The idea is that certain animals are capable of mindreading, mindreading is uniquely conscious and, therefore, the animals are conscious. In light of the findings described above, however, this argument doesn't go through; the second premise is simply false. This means that we can't move from evidence that animals can mindread to the conclusion that they are conscious.

state of the universe, the laws of nature, and the outcomes of random quantum mechanical events, everything that happens is determined by what came before it. If something like this picture is right, no one could ever act in a way other than the way he does act. So, unless we're willing to concede that human beings never have control over – or are responsible for – their thoughts and behaviors, we can't cash out control as *regulative* control.

According to the above challenge, the problem with regulative control is that it isn't humanly possible. Given the way the world works, human beings can't exert regulative control over their thoughts and behaviors. We want to preserve the possibility of human control. Therefore, we shouldn't characterize control as regulative control. Frankfurt (1971) offers a different sort of challenge. According to Frankfurt, regulative control would be problematic even if it weren't practically impossible; even on its own merits, it's the wrong way to think about control. To illustrate this, he encourages us to imagine that a puppetmaster, Black, has managed to gain access to the brain of another individual, Jones. If Jones does what Black wants him to, Black won't interfere in his neural processes but, if he tries to do something else, he will intervene. Suppose that Jones does what Black wants him to do without any intervention from Black. Though Jones couldn't have acted other than he did act, Frankfurt contends, he still acted freely. Though Black *would have* prevented him from acting against his wishes, this possibility doesn't strip Jones of control over the action he actually did perform.

These challenges suggest that what distinguishes controlled from uncontrolled behaviors (or behaviors for which we are and aren't responsible) is not the space of alternative possibilities. Rather, the difference seems to consist in something about the

way in which a behavior comes about. What differentiates a controlled behavior from an uncontrolled behavior is something about *the nature of the process that produces it*. What is this something? One possibility is that controlled behaviors are accompanied by a *feeling* or *experience* of control. On this account, controlled behaviors are behaviors that *feel* controlled.

The trouble with this account of control is that an *experience* of control over (or *willing* of) a behavior doesn't seem to be either necessary or sufficient for *actual* control over the behavior. Behaviors that we would intuitively classify as controlled can occur in the absence of feelings of control, and behaviors we wouldn't intuitively classify as controlled can feel controlled. For an example of the first phenomenon, consider patients with alien hand syndrome. These patients feel disconnected from their hands; they feel as if their hands are controlled by external forces. Despite the fact that they don't feel as if they are moving their hands, however, they *are* moving them. Similarly, table turners and Ouija practitioners don't feel as if they are moving tables or Ouija pointers, respectively, but they *are* moving them.

Wegner & Wheatley (1999) provide evidence of the second type of case. They paired participants with experimental confederates, and asked both participants and confederates to put their hands on a computer mouse and use it to move a cursor around a screen. At particular intervals, the participant-confederate pairs were instructed to pick a time to stop moving the cursor. In some cases, the confederate determined when the cursor would stop and, in others, he allowed the participant to do so. Wegner & Wheatley found that, under certain conditions, participants felt as if they had decided when to stop the cursor even when the confederate was actually in control.

If neither the space of alternative possibilities nor phenomenology marks the difference between controlled and uncontrolled behaviors, what does? What has to be true of a process for the behaviors it produces to qualify as controlled? To answer this question, let's turn back to the Bateman and RoboBateman scenarios. What are some differences between Bateman's behavior and RoboBateman's? Unlike RoboBateman, Bateman weighs the pros and cons of a range of possible responses to Allen's approach – doing nothing, exchanging pleasant hellos, tripping Allen, etc. – and decides on the tripping response. This prompts him to form the intention to trip Allen, which leads him to initiate the relevant motor processes. Throughout, he can monitor his states and processes, checking whether the intention on which he acts meshes with his other desires and whether the processes he initiates remain on target to achieve his intentions. If these processes seem to be veering off track, he can intervene to correct them. RoboBateman engages in none of this processing. He isn't even aware that Allen is approaching, so he certainly doesn't deliberate about a response to his approach, form an intention to act in any particular way, or check this intention against his other commitments. He also lacks the ability to monitor or correct the process that produces his leg spasm. This process is *ballistic*; once initiated, it can't be revised or stopped.

These differences suggest two general contenders for the relevant notion of control:

(1) *Deliberative control*: a behavior is controlled provided that it follows from a deliberative process.

(2) *Metacognitive control*: a behavior is controlled provided that the behaver has higher-order oversight over the process that produces it.

Let's spell out each of these ways of thinking about control in a bit more detail.

First, deliberation consists in reflection on reasons. Cashing out control in terms of this kind of reflection is a fairly common move. According to a popular philosophical understanding of control – attributed by Suhler & Churchland (2009) to Doris (2002), among others, and referred to as neo-Kantianism – controlled behaviors are reflective behaviors. Fischer (1987) seems to endorse a similar view. According to his reasons-responsiveness account of control, a behavior is controlled provided that the mechanism that produces it is responsive to a sufficiently wide range of reasons; in possible worlds in which there is sufficient reason for it to produce a different behavior than it produces in the actual world, the mechanism recognizes this reason as sufficient and acts on it. This suggests that, for Fischer, control consists in reflecting (in a particular way) on available reasons.

Second, there are two main ways to cash out metacognitive accounts of control. According to the first, a behavior is controlled provided that the first-order intention, or desire, that leads to the behavior is higher-order endorsed by the behaver. Frankfurt's (1971) hierarchical mesh theory is an example of this kind of account. Frankfurt asks us to consider three drug addicts. All three addicts have a first-order desire to use drugs, but one (the willing addict) also has a second-order desire to desire to use drugs, the second (the unwilling addict) has a second-order desire not to desire to use drugs, and the third (the wanton addict) has no second-order desires about his desire to use drugs. According to the hierarchical mesh theory, the first of these addicts has control over his behavior while the second and third don't. The reason for this is that there is the right kind of *mesh* between the willing addict's first- and second-order desires, but not the unwilling or wanton addicts'. The willing addict, but not the unwilling addict or the wanton addict,

higher-order endorses his first-order desire to use drugs. Why does Frankfurt think it's important to have this mesh between desires? The basic gist is that controlled behaviors are behaviors that reflect the behaver's real self, and higher-order-endorsed behaviors are more likely to reflect this real self than unendorsed behaviors. They are more likely to be expressions of the behaver's settled commitments.

The second way of cashing out metacognitive control is in terms of oversight over the mechanism of the process that produces a thought or behavior. According to this interpretation of metacognitive control, a behavior is controlled provided that the behaver can monitor and, if necessary, correct the processes that produce the behavior. This is the understanding of control championed by Hassin (2005). The easiest way to illustrate it is by contrasting it with reflexive behavior. Suppose I tap you on the knee, right at the intersection of your femur and tibia. This will cause your lower leg to snap up. If you're like most people, you won't have any access to the motor processes that lead to this reflexive leg movement, and you won't be able to intervene in them. Once your knee has been tapped (and barring external intervention), your reflexive motor processes will run to completion. By contrast, behaviors that are controlled in Hassin's sense can be monitored and corrected. We can check the progress of these processes, and intervene in them if they require correction. We can stop or revise them, rather than simply waiting for them to play out.

Is there reason to think either of these kinds of control might be necessarily conscious? According to Suhler & Churchland's (2009) neo-Kantians, deliberative control is a thoroughly conscious affair. The primary inputs to deliberations are reasons and, for something to be a genuine *reason* rather than a mere *cause*, it has to be

conscious. Also, deliberation involves conscious reflection on reasons. In order genuinely to deliberate about something, you have to weigh relevant considerations against each other *consciously*. At least according to the neo-Kantian, then, consciousness is both input and mechanism constitutive necessary for deliberative control.

What about metacognitive control? There is a *prima facie* plausibility to the claim that metacognitive control is necessarily conscious. Suppose I try to sell you a device. I describe this device as a baby monitor, but it doesn't have either a video or an audio feed. Indeed, the device will provide you with no sensory information about your baby's well-being whatsoever. You're unlikely to be sold on this device. And you might be similarly unsold on the possibility of *cognitive* monitoring in the absence of phenomenological feedback. It might seem similarly improbable that we could monitor or correct mental states or processes without introspective access to those states or processes.

So, there are arguments that both of the main contenders for the kind of control involved in responsibility require consciousness. Do these arguments go through? Is consciousness necessary for deliberative control or either type of metacognitive control? According to the neo-Kantian tradition Suhler & Churchland (2009) cite, reasons have to be conscious. Is this true? Or can there be unconscious reasons? To answer this question, we need to get a clearer sense of what it means for something to be a reason. What distinguishes genuine *reasons* for behavior from mere *causes* of behavior? My current typing behavior is caused by a particular cascade of neural activations, a desire to finish this chapter, etc. Which (if any) of these causes are reasons? And why? What makes some of the causes of my behavior genuine reasons and the others nothing more than causes?

Suppose I am running down the street. We can explain this behavior mechanistically, in terms of the physics underlying my speed and direction, the neural events involved in my motor processes, etc. However, to get a *complete* picture of my behavior, we need to know something else: what's the purpose of my running? Am I trying to complete a marathon? Am I chasing a wallet snatcher? Am *I* the wallet snatcher? Like the mechanistic facts, these are (possible) causes of my behavior. Unlike those facts, though, they aren't *mere* causes; they're also (possible) *reasons*. As the questions we're asking suggest, what distinguishes these reasons from mere causes is *purposiveness*. Reasons are possible causes of behavior that are directed toward a purpose. If the purpose of my running down the street is to complete a marathon, for example, marathon-completion is my reason for running down the street.

If this is right, we can rephrase the question about unconscious reasons as follows: can there be unconscious purpose-directed causes of behavior? There seems, to me, to be no reason whatsoever to think that purpose-directed causes are necessarily conscious. Indeed, as Matthews (2005) points out, there's actually reason to think they *aren't*. Take, for example, the case – adapted from Ryle (1949) – of the good tennis player. The ways the good tennis player positions her racket, moves her feet, etc. are all directed at purposes. However, given the speed of play and the limitations of conscious processing, she can't work through all the relevant considerations consciously. She acts for reasons, but these reasons aren't all – and, indeed, *couldn't* all be – *conscious* reasons. This case isn't unique. As Matthews points out, there are many other cases of similarly unconscious purpose-directed causes. To maintain that reasons are necessarily conscious is to fail to take account of a range of clear, common, everyday examples of unconscious reasons.

In response to this, the proponent of necessarily conscious reasons might simply *stipulate* that the only things that merit the label 'reason' are conscious purpose-directed causes. Though there can be unconscious causes (and even unconscious purpose-directed causes) of behavior, none of these causes rises to the level of genuine reasons. Now, there are a couple of different ways to interpret this claim. One is as purely semantic. On this level, the claim isn't very interesting. In and of themselves, battles over who gets to use a particular word are rarely worth fighting. The other way to interpret the claim, on the other hand, gets at something deeper. According to this interpretation, the claim is that, no matter what we *call* them, unconscious causes – even unconscious purpose-directed causes – can't play the role genuine reasons play in deliberative processes.

This second interpretation of the claim is much more interesting than the first. However, there's little reason to think it's true. Deliberations are weighings of considerations for and against an action or range of actions. The role reasons play in this kind of processes is the considerations role; reasons are the considerations that are weighed against each other. For the second interpretation of the claim to be true, it would have to be the case that unconscious considerations can't be weighed against each other. But this *isn't* the case. As we saw in the tennis player example, unconscious considerations enter into the processes that produce foot, racket, etc. movements. In those processes, various unconscious considerations are weighed against each other to produce decisions about where to place a foot, how to tilt the racket head, etc.

What about deliberation itself? Can we unconsciously weigh considerations against each other? Relatedly, can unconscious mechanisms be reasons-responsive? Can unconscious mechanisms take a sufficiently wide range of reasons into account to satisfy

Fischer's (1987) requirements? A 2004 study by Dijksterhuis gives us reason to think they can. Participants in this study were presented with a range of options, and instructed to choose between them. Some participants were told to make their decisions immediately after hearing the task instructions (immediate decision condition) while others were asked to engage in a few minutes of conscious thought about the task first (conscious thought condition), and still others were first given a few minute-long distracter task (unconscious thought condition). Dijksterhuis found that the decisions participants in the unconscious thought condition made were better – both when measured objectively (against pilot study findings about the importance of particular decision criteria) and when measured subjectively (against the participants' own judgments about the importance of criteria) – than those made by participants in the other conditions. These findings suggest that the unconscious thought participants engaged in very effective *unconscious* decision-making.

Effective decision-making is a paradigmatically deliberative process; to make a good decision, we have to weigh considerations for and against the available options. More specifically, we have to weigh a *wide range* of considerations for and against the available options. No matter how good our reasoning processes are, we won't make good decisions unless we are reasoning with reasonably complete information. Now, by all accounts, the participants in the unconscious thought condition made good decisions. Presumably, therefore, they engaged in deliberation and, in the course of this deliberation, took into account a reasonably wide range of relevant considerations. And they did so *unconsciously*. If this interpretation of Dijksterhuis' (2004) findings is right, we're hard-pressed to deny that deliberation can occur unconsciously.

Having addressed deliberative control, let's turn to the second type of control – metacognitive control. Above, I noted that there's a *prima facie* plausibility to the claim that metacognitive control requires consciousness. Ultimately, however, this *prima facie* plausibility runs afoul of the empirical record. For one thing, it runs up against evidence of reverse priming. Typically, priming works in the following way: prime stimuli that are similar to target stimuli speed up (facilitate) responses to the target stimuli, and prime stimuli that are dissimilar to target stimuli slow down (inhibit) responses (Fazio et al., 1986). For example, suppose two different people are primed with the word 'chocolate.' One is then asked to say 'cake' while the other is asked to say 'spinach.' The first person will say his word (cake) much more quickly than the second person will say hers (spinach). Recently, however, Glaser & Banaji (1999) stumbled upon a puzzling exception to this typical effect. When testing for racial priming effects, they found that participants sometimes display the *opposite* of the typical priming effect. In these cases, prime stimuli that are *dissimilar* to the target stimuli facilitate responses to the target stimuli while *similar* primes inhibit them. Glaser & Banaji dubbed this effect 'reverse priming.'

Why does priming have the typical effect in some cases and the reverse effect in others? In most priming tasks, the prime stimuli are evaluatively *moderate*. They are words, images, etc. that don't elicit particularly strong evaluative reactions (e.g. 'potato'). The primes in reverse priming tasks, on the other hand, are evaluatively *extreme*. These words (e.g. 'tumor') do elicit strong evaluative reactions. Some primes – call them 'universally extreme primes' – are evaluatively extreme for most or all of the population. For example, most people have a strong negative reaction to the word 'tumor.' Other

primes – call them ‘locally extreme primes’ – are only evaluatively extreme for a subset of the population. For example, stereotype-related words (e.g. ‘shrew’) elicit a strong reaction in people who test high for chronic egalitarianism, but not people who test low. Universally extreme primes tend to elicit reverse priming effects in most people, irrespective of their group memberships (Glaser & Banaji, 1999). Locally extreme primes, on the other hand, elicit reverse priming effects primarily – or even solely – in members of the relevant subset of the population (Moskowitz et al., 1999).

What could account for these results? Why would participants respond differently to primes that evoked strong evaluative reactions than primes that evoked moderate reactions? The answer is that they are *overcorrecting*, or *overcompensating*, for the potential biasing effects of the extreme primes (Glaser & Banaji, 1999; Glaser & Kihlstrom, 2005). Priming with valenced words or images tends to bias later judgments about other valenced words or images. For example, priming with the positive word ‘chocolate’ biases subsequent judgments about the negative words ‘tumor’ and ‘shrew.’ When the potential for this kind of bias is salient to participants, they try to correct for it. In some cases, they take this correction too far. Rather than *neutralizing* the potential bias, they *reverse* it (Martin, 1986).

The reason participants display reverse priming effects *selectively* – in response to evaluatively extreme but not evaluatively moderate primes – is because the potential for bias is more salient with evaluatively extreme than evaluatively moderate primes. Extremity is a key determinant of prime salience; the more extreme a prime is, the more likely participants are to notice that it has the potential to bias their judgments (Herr et al., 1983). As we saw above, when participants are aware of the potential for bias, they tend

to try to correct for it. So, by increasing the likelihood that participants will notice the potential for bias, evaluative extremity also increases the likelihood that they will (over)correct for the bias. Reverse priming is more common in evaluatively extreme than evaluatively moderate cases because the potential for bias is more salient in evaluatively extreme cases.

We can think of what's happening in reverse priming cases in terms of a driving metaphor. Suppose you're a novice driver, driving at night for the first time. You get sleepy, and start drifting onto the right shoulder. If you're alone in the car, you might not notice that you're drifting until your car is head-first in a roadside ditch. Imagine, though, that your mother is in the car with you. As soon as you start to veer off-road, she alerts you – probably not very subtly – that you're drifting. In response, you quickly spin the wheel to the left. Unfortunately, due to your inexperience, you miscalculate how sharp a turn is required. Instead of getting squarely back in the original lane, you end up veering into the left lane. The biasing effect of prime words is like the off-road veering in this metaphor, and prime extremity is like your mother's voice in your ear (or her white-knuckled gripping of the dashboard). Like your mother's alert, the extremity of the prime words sends up a warning that you're veering off-track. And, like the novice driver, you try too hard to correct for the accidental detour. Instead of returning to a neutral midpoint, you end up displaying the *opposite* bias.

Importantly, the participants in reverse priming cases are not *consciously* correcting for potential biases. When debriefed, Glaser & Banaji's (1999) participants reported no conscious awareness of the correction processes they were employing. Indeed, they didn't even appear to be conscious of a *need* for correction; they didn't even

report conscious awareness of the *potential* for bias. Glaser & Kihlstrom (2005) also cite other reasons to doubt that participants were making corrections consciously. First, if participants were correcting for biases consciously, they would display a learning curve. They would exhibit typical priming effects in early trials, and only show reverse priming effects once they had developed a compensation strategy. This kind of picture is inconsistent with the data. Participants display a consistent pattern of typical priming in moderate conditions and reverse priming in extreme conditions; their responses don't change from early to later trials. Second, participants displayed this same consistent pattern even though they were unaware of a link between response times to targets and the evaluative (in)congruence of primes and targets. Almost *none* of the participants guessed that there was a link between response times and evaluative (in)congruence, but almost *all* of them displayed the pattern.[38]

How does the reverse priming evidence challenge the claim that consciousness is necessary for metacognitive control? The challenge to Hassin's (2005) version of metacognitive control is fairly obvious. According to Hassin, a behavior is controlled provided that it issues from a process that is monitored and, if necessary, corrected. The reverse priming evidence is most naturally interpreted as evidence for this kind of metacognition.

Does this evidence also bear on Frankfurt's (1971) variety of metacognition? I think so. More specifically, the locally extreme reverse priming evidence seems relevant. Recall the driving intuition behind Frankfurt's account: controlled behaviors are behaviors that reflect the behaver's real self, or settled commitments. In the locally

[38] Further evidence for unconscious metacognition comes from a study by De Neys et al. (2008). Using problems modeled on Kahneman & Tversky's (1973) famous base rate neglect problems, they show that participants unconsciously detect and resolve conflicts between the processes in dual process pairs.

extreme reverse priming cases, the participants who display reverse priming effects are the ones who test high for chronic egalitarianism. These participants reject responses that are driven by inegalitarian impulses, and attempt to modify them to responses that are more egalitarian. Put slightly differently, they reject responses that aren't accurate reflections of their deeper commitments, and would endorse responses that more accurately reflected those commitments. Given that these rejections and endorsements occurred unconsciously, this hints at a general capacity for unconsciously endorsing and rejecting motives that do and don't, respectively, reflect more deeply-held commitments. This can be interpreted as a capacity for Frankfurt-style control over behaviors.

4.4 Rationality

'Rationality' is an ambiguous term. Sometimes, when we talk about rationality, we're talking about the relationships between our mental states and behaviors. Suppose, for example, that I tell you my car has been stolen, stripped for parts, and cubed at the local junkyard, but I spend hours each day combing my neighborhood for it. Witnessing this, you might – quite reasonably – think of me as irrational. After all, the belief I verbally express (that my car is irretrievably gone) and the belief signaled by my behavior (that it's not) contradict each other. Call the kind of rationality at issue here 'formal rationality.' For a being to be formally rational, his mental states and behaviors must cohere with each other.

Another way to think about rationality is in terms of the nature of the processes that produce thoughts and behaviors. Suppose, for example, that I'm on an apartment hunt. Given a list of possible options, I arbitrarily pick one. I don't deliberate about my choice; it doesn't follow from any (even the most superficial and cursory) consideration

of reasons. Instead, I just randomly pick one. Here, again, there's something irrational about the way I behave. My irrationality in this case isn't, however, the incoherence of formal irrationality; there's nothing *incoherent* about making a snap judgment about an apartment. Rather, the irrationality stems from a failure to arrive at mental states and behaviors in the right kind of way. My apartment choice is irrational because it doesn't follow from a deliberative process. Call the kind of rationality at issue here 'procedural rationality.' For a mental state or behavior to be procedurally rational, it must follow from deliberation.

Is there reason to think that either of these types of rationality is necessarily conscious? Let's start with formal rationality. Suppose that I place three objects – a car, a bike, and a sandwich – in front of you, and ask you to determine which one of them is not like the other two. Generally, this would be an easy task. Suppose, though, that I impose the following constraint on you: you are not allowed to look at, touch, or otherwise perceptually interact with any of the objects. With this constraint, what was a simple task seems to become a pretty impossible one. McGinn (1982) thinks something similar is true of formal rationality. If we are able to introspect our mental states, we can determine which (if any) of them is inconsistent with the others. If we don't have introspective access to our mental states, on the other hand, detecting inconsistencies in them is no longer nearly so straightforward. Just as a lack of perceptual access makes it difficult to determine which of a group of *objects* is unlike the others, a lack of introspective access makes it difficult to determine which (if any) of our *mental states* is the odd man out. According to McGinn, then, consciousness is necessary for formal rationality because it

is what enables us to detect – and, in turn, take steps to resolve – mental inconsistencies.[39]

Is this really true, though? Do we have to be able to introspect our mental states in order to regulate them? Moran (2001) offers some reasons to doubt it. For one thing, we receive a continuous stream of new perceptual information and, to keep up with this stream, we have to constantly update our beliefs. This mental state regulation doesn't require conscious monitoring. Indeed, as Moran points out, it *couldn't* require conscious monitoring. Our conscious processing capacity is quite severely limited, especially relative to the amount of perceptual information we receive (and consequent belief updating we have to perform) on a moment-to-moment basis. If we had to monitor and respond to all of this information consciously, we wouldn't be able to function in the effective, efficient way we do. For another thing, if all mental state regulation required conscious monitoring, we'd quickly get tangled up in an infinite regress. Our lowest-level beliefs would have to be consciously monitored by second-level beliefs, which would have to be consciously monitored by third-level beliefs, and so on. Such an infinite regress isn't practically possible.

Now, it might be noted that Moran's (2001) points here only get us so far. They show that mental state regulation as a whole can't always require consciousness, but don't make the more specific point that *inconsistency detection* can occur unconsciously. What can we say here? Can inconsistency detection in particular happen unconsciously? The first thing to note is that unconscious inconsistency detection isn't as obviously challenged by McGinn's (1982) argument as it might initially seem. Awareness of mental

[39] Note that there are similarities between McGinn's (1982) argument here and the argument for the necessity of consciousness for cognitive monitoring and correction in the previous section.

states almost certainly is a prerequisite for detection of inconsistencies between them. However, this only commits us to the conclusion that consciousness is necessary for inconsistency detection if we equate awareness with consciousness (i.e. we take all awareness to be *conscious* awareness). And *equating* awareness and consciousness is too strong. Why? If you search for 'blindsight' on YouTube, you'll come across a rather striking video. In this video, a man walks the length of a hallway that has been 'booby trapped' with various obstacles. The man edges around each obstacle on his way from one end of the hallway to the other. What makes the video striking is that the man is cortically *blind*; he has no conscious visual experience. Because he has no conscious visual experience, he isn't *consciously* aware of the chairs and boxes in his path. If you asked him if he could see them, for example, he would say no. However, it would be a mistake to say that the man is *un*aware of the obstacles. After all, if he were unaware of them, how could he so neatly sidestep them?

So, unconscious awareness of mental states isn't ruled out by definition. But does it ever actually happen? Do we ever actually unconsciously access our mental states? Some evidence that we do comes from the literature on unconscious goal pursuit. Successful goal pursuit requires the goal pursuer to monitor his states to ensure that he remains on target to achieve his goal. As we'll see in Chapter 5, there's ample evidence that we can successfully pursue goals unconsciously.

There's a lingering further question, though. Formal rationality doesn't involve just unconscious *access* to mental states; it also involves *using this access to resolve inconsistencies*. To be formally rational, according to McGinn (1982), is not just to be aware of our mental states, but to resolve any inconsistencies we might find in them. Can

we do this kind of thing unconsciously? The evidence cited at the end of the previous section suggests that we can. Think, for example, about the locally extreme reverse priming cases. In those cases, individuals who tested high for chronic egalitarianism tended to (over)correct for potential non-egalitarian biases. For example, if primed with a stereotype word, they tended to show the reverse of normal priming effects. This suggests that, when they noticed that they were likely to give a response that clashed with their general egalitarian bent, they attempted to correct the response. Put another way, they seem to have tried to stave off potential inconsistency between their egalitarian beliefs and their primed responses. And they seem to have done so unconsciously.

Next, let's turn to the second kind of rationality – procedural rationality. Why might we think consciousness is necessary for this kind of rationality? To answer this question, we first have to have a better sense of exactly what procedural rationality involves. I think Rovane (1998) offers a good encapsulation of the intuitive understanding of procedural rationality. On her account, a mental state or behavior is procedurally rational provided that it follows from (attempted) consideration of everything in its possessor's rational point of view. I am procedurally rational provided that, when I'm making a decision, I at least attempt to account for all considerations that might be relevant to that decision.

Is there reason to think that this kind of rationality has to be conscious? One possible move to make here would be to stipulate that *rational* points of view are *conscious* points of view. However, this seems *ad hoc*. It's also at odds with the spirit of Rovane's (1998) project. One of Rovane's main goals is to develop an account of rationality that allows for the possibility of rational group agents; she thinks it's possible

that a group of human beings could come together to form a group that behaved procedurally rationally, and that our understanding of rationality should reflect this possibility. Now, group agents clearly don't have conscious points of view; they can't introspect what the group as a whole is thinking, planning, etc. So, defining rational points of view as *conscious* points of view would undermine one of Rovane's projects. Simply stipulating that rational points of view are conscious points of view isn't a good way to introduce consciousness into her version of procedural rationality.

However, there is another way in which consciousness might seem to be necessary for Rovane's (1998) type of procedural rationality: via mental simulation. Rovane says that all-things-considered judgments should take into account not just synchronic considerations, but also diachronic considerations; they should accommodate not only present, but also past and future events. In order to take past or future events into account, we have to be able to project ourselves into the past and future, and both past projection (episodic remembering) and future projection involve mental simulation (Goldman & Shanton, forthcoming; Shanton & Goldman, 2010). Therefore, for Rovane, procedural rationality seems to require simulation.

Now, as we saw in §4.2, Baumeister & Masicampo (2010) think simulation is necessarily conscious. It seems plausible that simulation is necessary for maintaining consistency between mental states over time. Therefore, if consciousness really is necessary for simulation, it's also necessary for Rovane's (1998) kind of procedural rationality. Of course, though, consciousness *isn't* necessary for simulation. As we also saw in §4.2, there's evidence that we can simulate unconsciously. In that section, we focused primarily on simulationist mindreading. However, there's also evidence that we

can engage in a range of other kinds of simulationist cognition, including consequentialist decision-making (Dijksterhuis, 2004), forming and pursuing goals (Bargh et al., 2001), and episodically remembering (DeCoster et al., 2006). Like formal rationality, procedural rationality seems to be something that can be achieved unconsciously.

4.5 Conclusion

There are many functions that are highly correlated with personhood. Some of these functions are clearly not necessary for personhood. For example, all persons (that we know of) can digest food, but food digestion isn't necessary for personhood. A being that got its energy from sunlight or electrical outlets could be a person. Other personhood-correlated functions, on the other hand, *do* seem necessary for personhood. For example, there's a case to be made that the correlation between rationality or language-like thought and personhood isn't merely accidental; these kinds of functions are plausibly *necessary* for personhood. Being necessary for personhood is one of the criteria that must be satisfied by a function for it to be the function of consciousness, so these functions are halfway to qualifying as that function. To *fully* qualify, however, they also have to satisfy the other criterion of the function of consciousness; they also have to be *unique* to consciousness. In this chapter, I tested local candidates for the function of consciousness against this criterion; I asked whether consciousness is necessary for any of them. The answer I came to in each case was no.

This conclusion is clearest in the case of being an object of the intentional stance in Dennett's (1976) sense. All that's required to be an object of the intentional stance in this sense is performing behaviors that can be explained in terms of intentional states. Infants can satisfy this criterion. So can dogs and plants and thermostats and geometric

shapes. If a *geometric shape* can be an object of the intentional stance in Dennett's sense, though, consciousness obviously isn't required. The same conclusion can also be drawn – if not quite as quickly or easily – about the stronger version of being an object of the intentional stance and being a subject of the intentional stance. Intentional states needn't be conscious, so unconscious beings aren't disqualified from possessing – or acting from – intentional states. Also, despite claims to the contrary by theorists like Humphrey (1983) and Baumeister & Masicampo (2010), consciousness isn't required for mindreading. We can explain behaviors in terms of mental states without ever consciously connecting the states to the behaviors, mindread without being conscious of the processes we use to do so, and even attribute mental states to stimuli of which we're consciously unaware. So, consciousness isn't necessary for mindreading – or being a subject of the intentional stance – in either a prerequisite, mechanism constitutive, or input constitutive sense.

Beings can also be responsible for their thoughts and behaviors without being conscious. The core of responsibility is control; we are responsible for our thoughts and behaviors when we have control over them. As pointed out above, there are a couple of different ways to interpret 'control' in this context. No matter which of them we choose, though, control – and, consequently, responsibility – isn't limited to conscious beings. First, reasons and deliberation about reasons can both be unconscious. So, if we think of control as deliberative – if what it means for a thought or behavior to be controlled is for it to be the product of a deliberative process – control can be unconscious. Second, metacognitive control can also be unconscious. Whether the metacognition occurs at the

input stage (as suggested by Frankfurt, 1971) or the mechanism stage (as suggested by Hassin, 2005), it can be deployed unconsciously.

The same type of conclusion holds in the case of rationality. As with 'control,' there's more than one way to cash out 'rationality.' Also as with control, though, it doesn't matter which of them we choose. On either understanding, unconscious beings can be rational. If we think of rationality formally – as coherence or consistency between mental states and behaviors – it's not limited to conscious beings. Not only can we be unconsciously *aware* of potential inconsistencies in our mental psychologies, but we can unconsciously intervene to *correct* them. Similarly, unconscious beings can be procedurally rational. We can unconsciously weigh reasons against each other, and choose thoughts or behaviors on the basis of these weighings. This is true even if, to do so, we have to take into account events that happened in the past or might happen in the future. Though factoring such past and future happenings into our decision-making involves simulation, that's no obstacle. As amply demonstrated by multiple studies of multiple simulationist tasks, we can simulate unconsciously.

Finally, the remaining two local candidates for the function of consciousness – language-like thought and self-monitoring – can also be performed unconsciously. This comes out in the course of investigating the other four local candidates. Simulation requires language-like thought – in order to simulate past, future, or hypothetical me or past, present, future, or hypothetical you, I have to be able to string events into sequences and combine them with each other – and mindreading and procedural rationality require simulation. Therefore, evidence that we can unconsciously mindread or be unconsciously procedurally rational is evidence that we can engage in unconscious language-like

thought. A similar chain of reasoning yields the conclusion that self-monitoring can be unconscious. We have to be able to monitor our mental states and processes to deploy metacognitive control, so evidence that we can unconsciously deploy metacognitive control is evidence for unconscious self-monitoring. Combined with the other conclusions reviewed in this section, this confirms that none of the local candidates satisfies our second criterion. None of them qualifies as the function of consciousness.

CHAPTER 5: GLOBAL CANDIDATES FOR THE FUNCTION OF CONSCIOUSNESS

5.1 The Global Workspace Theory

Picture a theater. There's a stage at the front of the theater, and a sea of audience members in the house. Actors crowd into the wings, waiting for their cues to go on. Once onstage, some of the actors are illuminated by a spotlight. According to an influential picture of the mind – Baars' (1988) Global Workspace Theory (GWT) – the mind is like this theater. Specialized unconscious processors produce states that compete for access to working memory. Some of the states that make it into working memory – what Baars describes as the 'active elements' of working memory – are illuminated by the 'spotlight' of attention. These attended contents of working memory are conscious states. They are broadcast to the specialized unconscious processors, to be used as inputs to further processing.

Now imagine an interdisciplinary working group. A group of experts – say, a philosopher, a psychologist, and a neuroscientist – has assembled in a room. A cognitive scientific problem has been posted on a blackboard at the front of the room. Each expert adds his own specialized knowledge about the problem to the board, writing something new or revising or erasing something that's already been posted. By the time they have all made their contributions, a complete solution to the problem has emerged. Using the blackboard as a communal workspace, the experts have combined their resources to solve a complex, interdisciplinary puzzle. According to the GWT, the mind is also like this interdisciplinary work group. Attended working memory functions as a mental workspace, where multiple processing resources can be coordinated to solve complex,

interspecialization problems. The processes that perform these coordination processes are conscious processes.

Each of these metaphors emphasizes a different role for consciousness in the mind. The theater metaphor emphasizes a broadcasting role. Just as the spotlighted actors' performances are broadcast to the audience members, conscious mental states are broadcast to the specialized unconscious processors. The blackboard metaphor emphasizes a coordination role. The blackboard enables the cognitive scientists to pool their resources. Similarly, attended working memory provides a space where various specialized processors can come together to work on a single problem.

Now, according to Baars (1988), consciousness isn't just *capable* of performing these two functions; it's *necessary* for them. He offers a couple of arguments for this claim. First, it's widely believed that attention and working memory have to be conscious. As we've already seen, on Baars' (1988) picture, both attention and working memory play absolutely central roles in broadcasting and coordination; broadcast states are the attended contents of working memory, and attended working memory is the 'workspace' in which various parts of the mind are coordinated. If attention and working memory are necessarily conscious, then, broadcasting and coordination must also be conscious.

Second, there are certain kinds of processes that require broadcasting and / or coordination, and these processes seem only to operate consciously. Two of the primary examples GWT theorists cite in this context are novelty processing and cognitive flexibility. When we come across something we've never encountered before, we don't yet have a 'template' for dealing with it. This means that we can't simply delegate it to a

pre-established ‘expert’ (unconscious processor). Instead, we have to combine the resources of multiple unconscious processors. Imagine, for example, that you’ve never seen or heard of a mermaid before. The first time you encounter a mermaid, you won’t have the resources to process it directly. By combining your knowledge about fish with your knowledge about women, however, you can start to understand what this new creature is. Broadcasting the stimulus to your *existing* knowledge sources, and coordinating their responses to it, provides you with the resources to process the *new* stimulus.

Specialized processors have their advantages. Most obviously, they are very efficient. Because their responsibilities are closely circumscribed – each deals with only a limited set of stimuli and produces only a limited set of outputs – they can be very fast and very good at their jobs. The *problem* with specialized processors is that, by themselves, they are inflexible. They are very good at what they are programmed to do, but they can’t go beyond their programming. If they run into something for which they haven’t been trained, they’re stumped.

To introduce flexibility into a system of specialized processors requires broadcasting and coordination. To respond flexibly to a problem, you have to have various possible responses available to you. Often, you have to be able to combine and recombine these various options in new and different ways. Various possible responses to a problem will only be available if the problem has been made available (broadcast) to more than one problem-solving process, and these responses can only be combined or recombined if the processes can be coordinated with each other.

Both novelty processing and cognitive flexibility are closely correlated with consciousness. As many GWT theorists have observed, consciousness tends to predominate in novel situations (see, for example, Armstrong, 1981 and Metzinger, 2009). Imagine, for example, that you're driving to work. You've driven this route every day for the past twenty years, so you're operating entirely on autopilot. Now imagine that road crews recently started a construction project in your area, and there's a new detour on your route. When you arrive at the detour, you'll snap out of autopilot. Dealing with this new obstacle seems to require your *conscious* attention.

Cognitive flexibility also seems to flip on and off with consciousness. Penfield (1975) describes cases of patients with epilepsy who lose consciousness during *petit mal* seizures. During seizures, these patients can perform high-level tasks like playing the piano and driving home from work, but their behaviors take on an automatic flavor; the piano player can't play new songs, the commuter can't account for detours, etc. When they lose *consciousness*, these patients also lose the ability to respond *flexibly* to their surroundings.

Baars' (1988) arguments for the necessity of consciousness for broadcasting and coordination are empirical arguments, and they generate empirical predictions. First, they predict that attention and working memory never operate unconsciously. The first argument is predicated on the assumption that these two operations are necessarily conscious. If they aren't, we can't draw its conclusions about the necessity of consciousness for broadcasting and coordination. Second, they predict that tasks like novelty processing and cognitive flexibility are never performed unconsciously. Baars seems to be right that broadcasting and coordination are constitutively necessary for

novelty processing and cognitive flexibility; broadcasting and coordination do seem to be parts of the processes that process novelty and respond flexibly. If he's *also* right that broadcasting and coordination are necessarily conscious, at least some parts of novelty processing and cognitively flexible processes must be conscious; these kinds of processes can't fly entirely under the introspective radar.

An obvious way to check whether broadcasting and coordination satisfy the second criterion of the function of consciousness is to test these predictions. Do attention and working memory really have to be conscious? Are we really incapable of processing novelty or responding flexibly without consciousness? In each of §5.2 and §5.3, I test one of these predictions. I close, in §5.4, by drawing out some implications of the test results.

5.2 Attention and Working Memory

Baars' (1988) first prediction is that attention and working memory never operate unconsciously. On its face, this prediction seems pretty plausible. Imagine that you've just been given a phone number to call. You don't have anything to write with, so you have to hold the number in your mind long enough to dial it. Holding the phone number in your mind is a paradigmatic example of an exercise of working memory. It also seems to be a paradigmatically conscious activity. On further reflection, the original description I gave of your phone-number-remembering doesn't seem quite specific enough. You don't seem to be just holding the number in your *mind*; you seem to be holding it in *consciousness*. Attention seems even more obviously to be necessarily conscious than working memory. Indeed, it's difficult even to see what it could *mean* for attention to be unconscious. How could we attend to stimuli of which we were unconscious?

In some recent studies, however, researchers have dug a little deeper into attention and working memory phenomena. These studies reveal that, though they might be the norm, conscious attention and conscious working memory aren't the rule. It's certainly true that attention and working memory are *usually* conscious, but they don't *have* to be. Despite its plausibility, Baars' (1988) first prediction doesn't actually hold.

To see this, let's take a look at the studies, starting with studies of unconscious attention. 'Attention' can be broken down into two parts:

(1) Attention shifting

(2) Attention proper

Attention shifting is the act of *directing* attention to a particular location in space while attention proper is the actual act of *attending* to the location. Some theorists argue that certain types of attention shifts are necessarily conscious. For example, Pierson & Trout (unpublished) maintain that endogenous attention shifts only occur consciously.[40] However, GWT theorists tend to focus on attention proper. Whether or not we have to *direct* attention consciously, they argue, we do have to be conscious of *attending*.

Surprising though it might seem, however, this claim doesn't hold up. Some evidence against it comes from a study with blindsight patient, G.Y. (Kentridge et al., 2004). Experimenters presented a box to G.Y.'s blind field then projected a line, oriented either vertically or horizontally, on it. G.Y. was instructed to report, as quickly as possible, whether the target line was oriented horizontally or vertically. In some trials, an arrow pointing to the part of the box where the line would appear preceded the line's

[40] Using a paradigm called the Posner response time paradigm, Decaix et al. (2002) make a compelling case against this claim. Their study convincingly demonstrates that we can endogenously direct attention unconsciously.

appearance. Though G.Y. performed quite well on all trials, his performance showed marked improvement on these 'cued' trials.

What does this suggest? G.Y. was able preemptively to attend to the locations cued by the arrows. He was quicker at reporting the orientations of the cued target lines because – thanks to the cues – he was already attending to their locations when they appeared. Even though he wasn't *conscious* of the locations, he was still able to *attend* to them. This finding isn't unique to G.Y. It has also been replicated in non-patient populations (Kentridge et al., 2008).[41]

So, attention can be unconscious. What about working memory? Hassin (2005) conducted a study that was specifically designed to address this question. He projected circles onto a 24x18 matrix, one at a time, in sets of five. The locations of the circles in some of the sets (rule sets) were governed by rules. In other sets (broken rule sets), the locations of the first four circles were governed by rules, but the location of the fifth was not. In control sets, none of the locations were governed by rules. Participants were asked to report, as quickly as possible, whether each circle was empty or filled in.

Hassin (2005) found that participants' reports about the fifth circle in a set were fastest when the set was a rule set, slower when it was a control set, and slowest when it was a broken rule set. This suggests that participants were extracting and using the available rules for each set. Applying the rules in the rule and broken rule conditions enabled them to make predictions about the location of the fifth circle. With rule sets,

[41] Further confirmation that we can attend to unconscious content comes from the finding that participants attend to images of nudes even when those images are masked (Jiang et al., 2006). It's also supported by evidence that change-detection (the opposite of change-blindness) can occur in the absence of consciousness (Fernandez-Duque & Thornton, 2000). This is significant because one of the primary sources of evidence for a necessary connection between attention and consciousness is supposed to be that participants are blind to changes that occur in unconscious fields or with unconscious stimuli.

these predictions enabled them to locate – and report on – the fifth circle more quickly. With broken rule sets, on the other hand, the predictions were counterproductive; participants expected the fifth circle to appear in the location predicted by the rules and, when it didn't, they had to take extra time to find it. This slowed down their reporting.

There's some disagreement among working memory theorists about exactly how to model working memory, but they have converged on similar accounts of its functional characteristics. As Hassin notes, they tend to agree that working memory tasks involve:

> (1) Active maintenance of ordered information for relatively short periods of time
>
> (2) Context-relevant updating of information and goal-relevant computations involving active representations
>
> (3) Rapid biasing of (task relevant) cognitions and behaviors in the service of currently held goals
>
> (4) Some sort of resistance to interference (2005, p. 202).

By this account, the rule extraction and use in Hassin's study was a working memory task. To perform it successfully, participants had to maintain lists of the circles they had already seen, and update these lists as new circles appeared. They then had to use this information to quickly and accurately make reports about the circles.

Importantly, the rule extraction and use was also *unconscious*. Participants weren't explicitly instructed to look for rules and, in post-experimental debriefings, they tended to deny that they had extracted or used rules. Even more tellingly, in post-experimental tests, they failed to display explicit knowledge or understanding of the rules. The moral to draw from this is that consciousness isn't actually necessary for working memory. Like attention – and *pace* GWT theorists – working memory can operate unconsciously.

5.3 Novelty Processing and Cognitive Flexibility

According to Baars' (1988) second prediction, novelty processing has to occur consciously. Is this prediction borne out? Initial reports from the cognitive unconscious front seemed to suggest so. Early in the cognitive unconscious tradition, researchers discovered that we unconsciously evaluate almost everything with which we come into contact; from cars to chairs to the friendly barista at the neighborhood coffee shop, we form unconscious impressions of pretty much every person, place, and thing we encounter (Chen & Bargh, 1997; Fazio, 2001; Winter & Uleman, 1984). The sole exceptions to this rule seemed to be novel stimuli. Unconscious processors seemed limited to re-evaluating stimuli that had previously been consciously evaluated; they didn't seem capable of forming impressions of unfamiliar stimuli.

The explanation early cognitive unconscious researchers gave for this was much the same as Baars' (1988) argument against unconscious novelty processing: unconscious processors can only engage in automatized, mechanical processing. When we consciously evaluate a stimulus, we create and store a 'file' of the evaluation. If we re-encounter a stimulus we have previously consciously evaluated, unconscious attitude-formation processes can retrieve this file and mechanically redeploy it. If we encounter a stimulus we haven't previously consciously evaluated (i.e. a *novel* stimulus), on the other hand, there's no file for unconscious processors to redeploy. In such cases, unconscious processors are out of their depth.

More recent findings, however, tell a different story. Duckworth et al. (2001) weren't convinced that the mechanical redeployment picture of unconscious attitude-formation was accurate, so they set out to test it. They started by priming participants

with either real positively- and negatively-valenced words or nonsense words that had previously been classified (by a different group of participants) as positive or negative. They then asked them to pronounce valenced target words as quickly as possible.

As we'd expect, participants who were primed with real words displayed priming effects. When primed with a positively-valenced real word, for example, they pronounced positive target words more quickly than negative target words. More surprisingly, participants also displayed priming effects when primed with *nonsense* words. Specifically, they displayed *evaluative* priming effects. When primed with the negatively-classified nonsense word *gumok*, for example, they were slower to pronounce positive than negative target words. Priming with the positively-classified nonsense word *talir* had the opposite effect.

Evaluative priming effects occur only if the valence of the prime is processed. Priming relies on recognition of congruence between the prime stimulus and the target stimulus. If a participant doesn't recognize the valence of the prime, the target, or both, he won't see them as evaluatively congruent with each other – and he won't display an evaluative priming effect. For participants to display evaluative priming effects with the nonsense primes, then, they must have processed their valences. They must have *evaluated* the primes.

Now, the nonsense stimuli in Duckworth et al.'s (2001) study were specifically designed to be novel (they were not words the participants would have encountered before), so these evaluations couldn't have been mechanical redeployments of previously (consciously) stored evaluations. Instead, they must have been actual unconscious

evaluations of the stimuli. If this is right, novelty processing isn't limited to conscious processes; we can also process new stimuli unconsciously.[42]

Novelty processing isn't the only kind of process that Baars (1988) takes to be necessarily conscious. He also thinks cognitive flexibility requires consciousness. Here again, though, his claim seems to go against the grain of the empirical evidence. Perhaps the most compelling evidence for unconscious cognitive flexibility comes from the unconscious goal pursuit literature. To pursue a goal successfully, you have to be aware of – and able to adapt to – unexpected changes in circumstances. Suppose, for example, that I have the goal of writing a spec script for my favorite TV show. I carefully plan out a writing schedule, setting aside a certain number of hours each day to work on the script. Now suppose that I'm unexpectedly called on to teach an extra class. To achieve my spec script-writing goal, I have to adapt to this change in my circumstances. I have to recognize that there's been a change, and adjust my trajectory to account for it. Inability to recognize or adjust to the change in my workload will prevent me from achieving my goal. Successful goal pursuit is, therefore, a sign of cognitive flexibility.

As it turns out, there's a wealth of evidence that goals can be successfully pursued unconsciously. To get a sense for this literature, let's consider a sample study from it. Bargh et al. (2001) unconsciously primed participants with cooperation-related words

[42] Treisman (1964) offers further confirmation of this. Using a dichotic listening task, she showed that participants can unconsciously process novel combinations of words in unfamiliar passages. It's also supported by evidence that we can unconsciously solve novel problems [e.g. Berry & Broadbent's (1984) novel sugar factory simulation problem]. Indeed, we can even solve novel problems we don't have the *capacity* to solve consciously, like learning grammar and event covariations. Our conscious processing capacities are quite shockingly limited – most estimates place the rate at which we receive information at approximately 11,000,000 bits per second, and the rate at which we can consciously process it at approximately 40 bits per second (Wilson, 2002). Even if these estimates were off by many orders of magnitude, our conscious processing capacities would still be insufficient for handling the complex algorithms involved in learning grammar and co-variations between events (e.g. that lightning is followed after a predictable interval by thunder). Nonetheless, we *can* learn them. Indeed, even young children can learn them – and they do so with relative ease (Czyzewska et al., unpublished; Lewicki et al., 1992).

(e.g. ‘dependable,’ ‘fair,’ and ‘friendly’) then had them play a resources dilemma game. In resources dilemma games, all players share a common pool of resources. It’s in each player’s individual interest to use as many of the resources as possible but, if all players use as many resources as they can, they will quickly be consumed and nothing will be left for anyone. Therefore, resources dilemma game play involves a trade-off between individual and communal interests.

Players can play this kind of game either competitively (prioritizing their own individual interests) or cooperatively (prioritizing communal interests). Somewhat surprisingly, Bargh et al. (2001) found that participants who were unconsciously primed with cooperation-related words were significantly more likely than control participants to play resources dilemma games cooperatively. Even more surprisingly, they were just as likely to play cooperatively as participants who had been *explicitly instructed* to cooperate. This suggests that, like those participants, they formed and pursued a cooperation goal. Unlike those participants, though, they did so *unconsciously*.

Now, there are a couple of objections that might be raised to the conclusion that these participants were engaging in unconscious goal pursuit. First, it might be objected that they weren’t really pursuing goals *unconsciously*. We don’t only form intentions because people tell us to. There can be any number of reasons to behave cooperatively in a resources dilemma game. For example, you might be naturally inclined to cooperate with others when playing games, or you might think cooperating will benefit you in the long run. In either of these – and probably a host of other imaginable – scenarios, you will form a cooperation goal even if you haven’t explicitly been told to do so. Therefore,

the fact that the primed participants weren't explicitly instructed to cooperate doesn't preclude the possibility that they *consciously* formed and pursued cooperation goals.

Second, it might be objected that the primed participants weren't really engaged in *goal pursuit*. Other experiments have shown that perceptual priming can directly elicit prime-consistent behavior. For example, participants who are primed with professor-related stereotypes subsequently perform particularly well on trivia games (Dijksterhuis & van Knippenberg, 1998), and participants who are primed with rudeness-related words are subsequently particularly likely to interrupt others (Bargh et al., 1996). Something like this might be going on in the cooperation study. I've been interpreting Bargh et al.'s (2001) findings as support for the following story: priming with cooperation-related words leads to formation of a goal to cooperate, which results in cooperation-pursuant behavior. The perceptual priming study suggests a different story: cooperation priming leads *directly* to cooperation-pursuant behavior. On this second story, there's no intermediate goal-formation step; participants aren't forming or pursuing goals.

Ultimately, though, neither of these objections hits the mark. Let's start with the first objection. If the only reason to think that the unconscious thought participants weren't cooperating consciously was that they hadn't been explicitly instructed to cooperate, the first objection might have legs. However, this *isn't* the only reason. Bargh et al. (2001) also collected debriefing data. After the game was over, they asked participants whether they had intended to behave cooperatively. During debriefing, primed participants tended explicitly to deny that they had consciously formed or pursued cooperation goals. Combined with the lack of explicit instruction to cooperate, this leaves us with little reason to believe that they were cooperating consciously.

What about the second objection? Perceptual priming effects – like those in the trivia and rudeness studies – decay over time; the more time that elapses post-priming, the less likely participants are to display the effects (Anderson, 1983). Commitment to goal pursuit, on the other hand, increases in strength until the goal is achieved (Atkinson & Birch, 1970). In a follow-up to the study described above, Bargh et al. (2001) primed participants with achievement-related words then asked them to work on a word search. Some were told to begin the word search immediately after priming while others were first given a short distracter task. If the priming effects were *perceptual* priming effects, we would expect them to be less pronounced in the latter condition than the former. If they were genuine *goal* priming effects, on the other hand, we'd expect the reverse pattern. Bargh et al.'s findings support the second of these predictions. Participants in the delayed condition performed significantly better on the word search than participants in the immediate condition.

In addition to increasing in strength over time, goal pursuit has two other characteristic features:

(1) Persistence in the face of obstacles (Gollwitzer, 1990; Gollwitzer & Bargh, 1996)

(2) Resumption after interruption (Gollwitzer & Liu, 1995)

Participants who have been primed with goal-related words display both of these characteristic features. In two further follow-up studies, Bargh et al. (2001) again primed participants with achievement-related words and asked them to perform a word search. In the first study, participants were given a few minutes to work on the word search then told, via intercom, to stop working. Using a hidden camera, Bargh et al. discovered that

primed participants were significantly more likely than unprimed participants to continue working after time had been called.

In the second study, participants were told that they would be performing two tasks – the relatively boring word search and a much more appealing cartoon strip task – starting with the word search. After they had worked for a few minutes, they were interrupted by an (experimenter-controlled) equipment failure. Once the equipment had been 'repaired,' they were told that there was only time for one task and allowed to choose which to perform. Primed participants were significantly more likely than unprimed participants to choose to return to the (boring) word search.[43]

Combined, these conclusions and the conclusions in §5.2 make a compelling case against the claim that broadcasting and coordination are necessarily conscious. The current conclusions show that tasks for which broadcasting and coordination are constitutively necessary, like novelty processing and cognitive flexibility, can be performed unconsciously. The conclusions in §5.2, show that the essential elements of broadcasting and coordination – attention and working memory – can operate outside consciousness. Like the local candidates before them, then, the global candidates fail to satisfy our second criterion – or qualify as the function of consciousness.

5.4 Architectural Implications

As noted in Chapter 3, broadcasting and coordination are *centralized* functions. Broadcasting involves a centralized message center, which takes in messages from various processors and transmits them to other processors. Coordination involves a

[43] Additional evidence for unconscious cognitive flexibility comes from work with the Wisconsin Card Sorting Task and Iowa Gambling Task (Hassin et al., 2009), and on implicit attitude-formation (Ferguson & Bargh, 2002), behavioral mimicry (Bargh & Morsella, 2008), and implicit memory in patients with amnesia (Graf & Schacter, 1985; Schacter & Graf, 1986).

centralized workspace, where various processors can come together to work on big-picture problems. Therefore, the finding that broadcasting and coordination can occur unconsciously suggests that there can be unconscious centralization. There's a centralized information exchange that operates unconsciously.

Baars' (1988) model of the mind doesn't have space for such an unconscious information exchange. As I've noted, he thinks that unconscious processes are limited to circumscribed, specialized, localized processing, and centralization is the domain of consciousness. Therefore, we have to revise his model of the mind. We have to make space in it for the possibility of unconscious centralization.

This raises the interesting question of exactly *how* we should make such space. How should we revise Baars' (1988) picture to accommodate the evidence of unconscious centralization? There are two main options here:

(1) Posit a single theater / workspace with conscious and unconscious operating modes

(2) Posit two theaters / workspaces: one conscious and the other unconscious

There's a short passage in his book in which Baars briefly entertains the idea of unconscious broadcasting and coordination. In this passage, he seems to favor the first of the above options. However, there's reason to prefer the second. One thing GWT theorists generally agree on is that centralized information exchanges are limited in capacity; only one process can call on them at any given time. There's only space in the centralized theater or workspace for one thing to be broadcast or one set of things to be coordinated at a time, so only that one thing or set of things can be *processed* at a time. If we want to broadcast or coordinate something else, we have to bump the first thing out of the spotlight. A consequence of this picture is that, if there was just one information

exchange, we wouldn't be able to perform two broadcasting or coordination tasks at the same time.

As it turns out, however, we *can* perform two broadcasting or coordination tasks at the same time. Think, for example, about the dichotic listening study footnoted earlier (Treisman, 1964). Treisman's bilingual participants were able to process two different streams of novel combinations of words at once. Many of the other studies we've encountered in the foregoing chapters further confirm that we can perform two broadcasting or coordination tasks at the same time. For example, Perruchet's (1985) conditioning study shows that we can simultaneously engage in conscious and unconscious learning, and Sanfey et al.'s (2003) ultimatum game study shows that we can simultaneously engage in conscious and unconscious decision-making.

This suggests that the better revision of Baars' (1988) picture is the second one. Rather than a single theater / workspace with conscious and unconscious operating modes, there are two theaters / workspaces – one conscious and the other unconscious. If this is right, there are two centralized authorities in the mind. Not only is there a *conscious* centralized authority but, it seems, there's also an *unconscious* centralized authority.

CODA

In Chapters 1-2 of this work, I showed that human beings can't be single, unified persons. There are conscious and unconscious forces operating in each human being, and these forces compete for control over his thoughts and behaviors. The disparities between these forces – in terms of both the kinds of processes they employ and the kinds of outputs they generate – are just too great for them to be reconciled into a single, unified person. They too regularly, and too thoroughly, conflict with each other to be best classified as parts of one and the same person.

In Chapters 3-5, I investigated the nature of the unconscious force. Drawing on work in the cognitive unconscious literature, I built up a picture of the unconscious. According to this picture, the set of unconscious states and processes is capable of performing the functions that are the most plausible candidates for criteria of personhood. It also seems to contain a centralized unconscious information exchange ('theater' and 'workspace') that's analogous to the centralized conscious information exchange posited by GWT. Combined, these two points suggest that there's a centralized information exchange in the unconscious which is capable of taking the intentional stance and having that stance taken toward it, thinking rationally, bearing responsibility for the thoughts and behaviors it produces, engaging in language-like thought, and monitoring its own cognitive states and processes.

In Chapter 3, we saw that it's notoriously difficult to identify necessary conditions of personhood. Unfortunately, identifying sufficient conditions is not much easier (Dennett, 1976). However, it seems clear that the being I've just described – the *centralized, capable, high-level* unconscious – has a serious claim to personhood. Not

only can it perform the functions we most closely associate with persons, but it also has the kind of centralized architectural structure we expect persons to have. Of course, I'd need more arguments to establish beyond a doubt that the set of unconscious states and processes is a person. But the signs do seem to be pointing in that direction. The door is certainly open to the possibility that normal human beings contain two separate persons.

Whether or not we accept this more radical conclusion, the arguments in Chapters 1-5 still push us to make substantial revisions to some of our long-held beliefs about the mind. For one thing, we can no longer think of persons and human beings as coextensive. There are genuine divisions in the human mind and, whatever the nature of the parts – whether they are each persons in their own right or not – these divisions prevent us from thinking of human beings as single, unified persons. For another, we should stop thinking of certain functions as unique to consciousness, or drawing the line between persons and non-persons at the boundary of consciousness. Whether or not consciousness is special, the arguments in Chapters 3-5 suggest that it isn't *metaphysically* special. There don't seem to be the kinds of qualitative differences between consciously- and unconsciously-performable functions that would warrant restricting personhood to conscious beings.

These revisions have implications for a range of philosophical issues. For example, the first revision has implications for responsibility and diachronic personal identity. If there isn't a single, unified person in each human being, responsibility for the human being's thoughts and behaviors is more diffuse than we tend to think, and we should approach determinations of diachronic personal identity differently than we historically have. It also has implications for the way we calculate well-being. If human beings and persons don't overlap, we can't assign well-being to human beings as wholes;

what's good for a human being as a whole isn't necessarily what's good for all (or any) of the persons he contains. The second revision has implications for the kinds of evidence we can use to test for the presence of consciousness. For example, evidence that a bird or monkey or dog can engage in mindreading can't be taken as evidence that that bird or monkey or dog is conscious. The changes suggested by this work could, therefore, ramify widely through many areas of philosophy. They have implications not only for philosophers of mind but also philosophers in any other area – from metaphysics to ethics to epistemology – that relies on claims about the nature of the mind or the relationship between humans and persons.

BIBLIOGRAPHY

Anderson, J. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Armstrong, D. (1981). *The nature of mind*. Ithaca, NY: Cornell University Press.

Apperly, I. & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4), 953-70.

Atkinson, J. & Birch, D. (1970). *A dynamic theory of action*. New York: Wiley.

Baars, B. (1988). *A cognitive theory of consciousness*. New York: Cambridge University Press.

Baars, B. (1997). In the theatre of consciousness: Global Workspace Theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4), 292-309.

Banaji, M. & Greenwald, A. (1995). Implicit gender stereotyping in judgments of fame. *Journal of Personality and Social Psychology*, 68, 181-198.

Bargh, J., Chen, M. & Burrows, L. (1996). The automaticity of social behavior: Direct effects of trait concept and stereotype activation on action. *American Psychologist*, 54, 462-79.

Bargh, J., Gollwitzer, P.M., Lee-Chai, A., Barndollar, K. & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, 81(6), 1014-1027.

Bargh, J. & Morsella, E. (2008). The unconscious mind. *Perspectives on Psychological Science*, 3, 73-79.

Bauer, R. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the Guilty Knowledge Test. *Neuropsychologia*, 22(4), 457–69.

Baumeister, R., Bratslavsky, E., Muraven, M. & Tice, D. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74, 1252-1265.

Baumeister, R., & Masicampo, E. (2010). Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal-culture interface. *Psychological Review*, 117, 945-971.

Baumeister, R. & Sommer, K. (1997). Consciousness, free choice, and automaticity. In R. Wyer, Jr. (Ed.), *Advances in Social Cognition (Vol. X)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Berry, D. & Broadbent, D. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, 36A, 209-231.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.

Blumstein, S., Milberg, W. & Shrier, R. 1982. Semantic processing in aphasia: evidence from an auditory lexical decision task. *Brain and Language*, 17, 301-315.

Bogen, J. (1993). The callosal syndromes. In K. Heinemann & E. Valenstein (eds.), *Clinical neuropsychology*. New York: Oxford University Press.

Bräuer, J. Call, J., & Tomasello, M. (2004). Visual perspective-taking in dogs (*Canis familiaris*) in the presence of barriers. *Applied Animal Behaviour Science, 88,* 299-317.

Butler, A. & Cotterill, R. (2006). Mammalian and avian neuroanatomy and the question of *consciousness* in birds. *Biological Bulletin*, 211, 106–127

Call, J., Bräuer, J., Kaminski, J. & Tomasello, M. (2003). Domestic dogs are sensitive to the attentional state of humans. *Journal of Comparative Psychology, 117,* 257-263.

Carruthers, P. (2009). Simulation and the first-person. *Philosophical Studies,* 114(3), 467-75.

Carruthers, P. (2010). Introspection: divided and partly eliminated. *Philosophy and Phenomenological Research*, 80, 76-111.

Chen, M. & Bargh, J. (1997). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid stimuli. *Personality and Social Psychology Bulletin*, 25, 215-224.

Choi, Y., Gray, H. and Ambady, N. (2005). The glimpsed world: Unintended communication and unintended perception. In R.R. Hassin, J.S. Uleman & J.A. Bargh (eds.), *The new unconscious*. New York: Oxford University Press.

Clements, W. & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, 9, 377-395.

Coslett, H. (1986, October). Dissociation between reading of derivational and inflectional suffixes in two phonological dyslexias. Paper presented to the Academy of Aphasia, Nashville, TN.

Cranford, R. & Smith, D. (1987). Consciousness: The most critical moral (constitutional) standard for human personhood. *American Journal of Law & Medicine*, 13, 233-48.

Czyzewska, M, Hill, T, & Lewicki, P. (Manuscript). Acquisition of information about conditional relations between variables in preschool children.

Decaix, C., Siéroff, E. & Bartolomeo, P. (2002). How 'voluntary' is voluntary orienting of attention? *Cortex*, 38, 841-5.

De Gelder, B., Morris, J.S. & Dolan, R.J. (2005). Unconscious fear influences emotional awareness of faces and voices. *PNAS*, 102(51), 18682-7.

De Haan, E., Young, A., & Newcombe, F. (1987). Face recognition without awareness. *Cognitive Neuropsychology*, 4, 385-415.

De Neys, W. & Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. *Cognition*, 106, 1248-1299.

DeCoster, J., Banner, M., Smith, E. & Semin, G. (2006). On the inexplicability of the implicit: Differences in the information provided by implicit and explicit tests. *Social Cognition*, 24(1), 5-21.

Dennett, D. (1971). Intentional systems. *Journal of Philosophy*, 68, 87-106.

Dennett, D. (1976). Conditions of personhood. In A. Rorty (ed.), *The identities of persons*. Berkeley, CA: University of California Press.

Dennett, D. (1990). The myth of original intentionality. In K. Mohyeldin Said, W. Newton-Smith, R. Viale and K. Wilkes (eds.), *Modelling the mind*. Oxford: Clarendon Press.

Dijksterhuis, A. (2009). Automaticity and the unconscious. In S. Fiske, D. Gilbert and G. Lindzey (eds.), *Handbook of social psychology, Volume 1 (5th Edition)*. Hoboken, NJ: John Wiley & Sons.

Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality & Social Psychology*, 87(5), 586-8.

Dijksterhuis, A. & van Knippenberg, A. (1998). The relation between perception and behavior or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, 74, 865-77.

Dominus, S. (2011, May 25). Could conjoined twins share a mind? New York Times. Retrieved May 25, 2011 from http://www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-share-a-mind.html.

Doris, J. (2002). Lack of character. Cambridge: Cambridge University Press.

Duckworth, K., Bargh, J., Garcia, M. T. & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, 13, 513–519.

Dunlop, C. (2000). Searle's unconscious mind. *Philosophical Psychology*, 13(1), 123-126.

Edelman, D. & Seth, A. (2009). Animal consciousness: A synthetic approach. *Trends in Neurosciences*, 32(9), 476-84.

Evans, J., Barston, J. & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295-306.

Fazio, R.H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, 15, 115-141.

Fazio, R., Sanbonmatsu, D., Powell, M. & Kardes, F. (1986). On the automatic evaluation of attitudes. *Journal of Personality & Social Psychology*, 50, 229-38.

Ferguson, M. & Bargh, J. (2002). Sensitivity and flexibility: Exploring the knowledge function of automatic attitudes. In L.F. Barrett & P. Salovey (eds.), *The wisdom in feeling*. New York: Guilford.

Fernandez-Duque, D. & Thornton, I. (2000). Change detection without awareness: Do explicit reports underestimate the representation of change in the visual system? *Visual Cognition*, 7, 323-344.

Fischer, J. (1987). Responsiveness and moral responsibility. In Schoeman, F. (ed.), *Responsibility, character, and the emotions*. Cambridge: Cambridge University Press.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68, 5–20.

Frith, C. & Corcoran, R. (1996). Exploring theory of mind in people with schizophrenia. *Psychological Medicine*, 26, 521–530.

Gazzaniga, M. (2003, January). The when, where, what, and why of conscious experience. Paper presented at the meeting of the National Institute on the Teaching of Psychology, St. Petersburg Beach, FL.

Gendler, T. (2008). Alief and belief. *The Journal of Philosophy*, 105(10), 634-663.

Gertler, B. (2009). Introspection. In P. Wilken, T. Bayne & A. Cleeremans (eds.), *The Oxford Companion to Consciousness*. Oxford: Oxford University Press.

Geschwind, N. (1981). The perverseness of the right hemisphere. *Behavioral and Brain Sciences*, 4, 106-7.

Glaser, J. & Banaji, M. (1999). When fair is foul and foul is fair: Reverse priming in automatic evaluation. *Journal of Personality & Social Psychology*, 77, 669-87.

Glaser, J. & Kihlstrom, J. (2005). Compensatory automaticity: Unconscious volition is not an oxymoron. In R. Hassin, J. Uleman & J. Bargh (eds.), *The new unconscious*. New York: Oxford University Press.

Goel, V. & Dolan, R. (2003). Explaining modulation of reasoning by belief. *Cognition*, 87(1), B11-B22.

Goldman, A. (Manuscript). Can unconscious states be introspected? Rutgers University, Department of Philosophy.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York: Oxford University Press.

Goldman, A. & Shanton, K. (Forthcoming). The case for simulation theory. In *Handbook of Theory of Mind* (Eds.), A. Leslie & T. German.

Gollwitzer, P. (1990). Action phases and mind-sets. In E.T. Higgins & R.M. Sorrentino (eds.), *Handbook of motivation and cognition (Volume 2)*. New York: Guilford Press.

Gollwitzer, P., & Bargh, J. (eds.). (1996). *The psychology of action: Linking cognition and motivation to behavior*. New York: Guilford Press.

Gollwitzer, P. & Liu, C. (1995). Willpower. In J. Kuhl & H. Heckhausen (eds.), *Encyclopedia of psychology: Motivation, volition and action*. Göttingen, Germany: Hogrefe.

Graf, P. & Schacter, D. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 501-518.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science,* 293, 2105-2108.

Greenwald, A., & Banaji, M. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.

Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Groeger, J. (1984). Evidence of *unconscious* semantic processing from a forced-error situation. *British Journal of Psychology*, 75, 305-314.

Güth, W., Schmittberger, R. & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4), 367–388.

Haidt, J. (2001). The emotional dog and its rational tail. *Psychological Review*, 108, 814-34.

Haidt, J., Koller, S. & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.

Hassin, R. (2005). Nonconscious control and implicit working memory. In R. Hassin, J. Uleman & J. Bargh (eds.), *The new unconscious*. New York: Oxford University Press.

Hassin, R., Bargh, J. & Zimerman, S. (2009). Automatic and flexible: The case of nonconscious goal pursuit. *Social Cognition*, 27(1), 20-36.

Heider, F. & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57, 243-59.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehre,, E., Gintis, H. & McElreath, R. (2001). In search of *Homo economicus*: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91,73–78.

Herr, P., Sherman, S. & Fazio, R. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of Experimental Social Psychology*, 19, 323-40.

Humphrey, N. (1983). *Consciousness regained.* Oxford: Oxford University Press.

Hurlburt, R. (1990). *Sampling normal and schizophrenic inner experience*. New York: Plenum Press.

Jacoby, L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30(5), 513-541.

Jacoby, L., Begg, I. & Toth, J. (1997). In defense of functional independence: Violations of assumptions underlying the process dissociation procedure? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 484-495.

James, W. (1890/1981). *Principles of psychology* (Vol. 2). New York: Holt.

Jiang, Y., Costello, P., Fang, F., Huang, M. & He, S. (2006). A gender- and sexual orientation-dependent spatial attentional effect of invisible images. *PNAS*, 103(45), 17048-17052.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.

Kant, I. (1788/1997). *Critique of practical reason*. Cambridge: Cambridge University Press.

Kentridge, R., Heywood, C. & Weiskrantz, L. (2004) Spatial attention speeds discrimination without awareness in blindsight. *Neuropsychologia*, 42, 831–835

Kentridge, R., Nijboer, T. & Heywood, C. (2008). Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia*, 46(3), 864-9.

Kunst-Wilson, W. & Zajonc, R. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, 207, 557-558.

Lewicki, P., Hill, T., & Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist*, 47, 796-801.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529-566.

Lormand, E. (1996). Nonphenomenal consciousness. *Nous*, 30, 242-61.

Martin, L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality & Social Psychology*, 51, 493-504.

Matthews, E. (2005). Unconscious reasons. *Philosophy, Psychiatry, & Psychology*, 12(1), 55-7.

McClelland, D. (1980). Motive dispositions: The merits of operant and respondent measures. In L. Weeler (ed.), *Review of Personality and Social Psychology (Vol. 1)*. Beverly Hills, CA: Sage.

McClelland, D., Atkinson, J., Clark, R. & Lowell, E. (1953). The achievement motive. New York: Appleton-Century-Crofts.

McClelland, D., Koestner, R. & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96, 690-702.

McGinn, C. (1982). The character of mind. Oxford: Oxford University Press.

Merikle, P. & Daneman, M. (1998). Psychological investigations of unconscious perception. *Journal of Consciousness Studies*, 5(1), 5-18.

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton: Princeton University Press.

Moskowitz, G., Gollwitzer, P., Wasel, W. & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality & Social Psychology*, 77, 167-84.

Nagel, T. (1974). What Is it Like to Be a Bat? *Philosophical Review*, 83(4), 435-50.

Nelkin, N. (1993). The connection between intentionality and consciousness. In M. Davies and G.W. Humphreys (eds.), *Consciousness: Psychological and philosophical essays*. Oxford: Blackwell.

Neuroskeptic. (2009, October 5). Is Freud back in fashion? No. Retrieved October 5, 2009 from http://neuroskeptic.blogspot.com/2009/10/is-freud-back-in-fashion-no.html.

Nichols, S. & Grantham, T. (2000). Adaptive complexity and phenomenal consciousness. *Philosophy of Science, 67*, 648-70.

Nichols, S. & Stich, S. (2003). "How to read your own mind: A cognitive theory of self-consciousness." In Q. Smith and A. Jokic (eds.), *Consciousness: New philosophical essays.* Oxford: Oxford University Press.

Nisbett, R. & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-59.

Olson, E. (2003) An argument for animalism. In: Martin, R. and Barresi, J., (eds), *Blackwell readings in philosophy: Personal identity*. Oxford: Blackwell.

Onishi, K. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.

Oosterbeek, H., Sloof, R. & van de Kuilen, G. (2004). Differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7, 171–188.

Penfield, W. (1975). *The mystery of the mind: A critical study of consciousness and the human brain*. Princeton: Princeton University Press.

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *Pavlovian Journal of Biological Science, 20(*4), 163-170.

Pierson, L. & Trout, M. (Manuscript). What is consciousness for?

Pitt, D. (2004). The phenomenology of cognition or what *is it like to think that P*? *Philosophy and Phenomenological Research*, 69(1), 1–36.

Posner, M.. & Snyder, C. (1975). Facilitation and inhibition in the processing of signals. In P. Rabbitt & S. Dornic (eds.), *Attention and performance:, Vol. 5*. San Diego, CA: Academic Press.

Prince, M. (1906). *The dissociation of a personality*. New York: Longmans, Green & Co.

Puccetti, R. (1993). Mind with a double brain. *British Journal for the Philosophy of Science*, 44 (4), 675-92.

Rovane, C. (1998). *The bounds of agency: An essay in revisionary metaphysics*. Princeton: Princeton University Press.

Rozin, P., Millman, L. & Nemeroff, C. (1986). Operations of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, 50(4), 703-712.

Ruffman, T., Garnham, W., Import, A. & Connolly, D. (2001). Does eye gaze indicate knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, 80, 201-224.

Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.

Sanfey, A., Rilling, J., Aronson, J., Nystrom, L. & Cohen, J. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626), 1755–1758.

Satpute, A. & Lieberman, M. (2006). Integrating automatic and controlled processes into neurocognitive models of social cognition. *Brain Research*, 1079, 86-97.

Schacter, D. & Graf, P. (1986). Preserved learning in amnesic patients: Perspectives from research on direct priming. *Journal of Clinical and Experimental Neuropsychology*, 8, 727-743.

Schulz, A. (2010). Simulation, simplicity, and selection: An evolutionary perspective on high-level mindreading. *Philosophical Studies*, 152, 271-285.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: The MIT Press.

Shallice, T. & Saffran, E. (1986). Lexical processing in the absence of explicit word identification: Evidence from a letter-by-letter reader. *Cognitive Neuropsychology*, 3, 429-458.

Shanton, K. & Goldman, A. (2010). Simulation theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 1(4)*, 527-38.

Sperry, R. (1974). Lateral specialization in the surgically separated hemispheres. In F. Schmitt & F. Worden (eds.), *Neuroscience 3rd study program*. Cambridge, MA: MIT Press.

Sperry, R., Gazzaniga, M. & Bogen, J. (1969). Interhemispheric relationships: The neocortical commissures; syndromes of hemispheric disconnection. In P. Vinken & G. Bruyn (eds.), *Handbook of clinical neurology, Vol. 4*. North-Holland: Amsterdam.

Sprong, M., Schothorst, P., Vos, E., Hox, J. & Van Engeland, H. (2007). Theory of mind in schizophrenia: Meta-analysis. *British Journal of Psychiatry*, 191, 5-13.

Steffens, M. (2005). Implicit and explicit attitudes towards lesbians and gay men. Journal of Homosexuality, 49(2), 39-66.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247.

Suhler, C. & Churchland, P. (2009). Control: Conscious and otherwise. *Trends in Cognitive Sciences*, 13(8), 341-7.

Treisman, A. (1964) The effect of irrelevant material on the efficiency of selective listening. *American Journal of Psychology*, 77, 533-546.

Uleman, J. (2005). Introduction: Becoming aware of the new unconscious. In R. Hassin, J. Uleman & J. Bargh (eds.), *The new unconscious*. New York: Oxford University Press.

Uleman, J. (2005). Introduction: Becoming aware of the new unconscious. In R. Hassin, J. Uleman & J. Bargh (eds.), *The new unconscious*. New York: Oxford University Press.

Van Baaren, R. (1999). A critical evaluation of Searle's Connection Principle. *Teorema*, 18(1), 73-83.

Warren, M. (1973). On the moral and legal status of abortion. *Monist*, 57(1), 43-61.

Wason, P. (1966). Reasoning. In B.M. Foss (ed.), *New horizons in psychology*. Harmondsworth: Penguin Books.

Wason, P. & Evans, J. (1975). Dual processes in reasoning? *Cognition*, 3, 141-54.

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

Wegner, D. & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, 480-492.

Weiskrantz, L. (1986). Blindsight: A case study and implications. Oxford: Oxford University Press.

Weiskrantz, L., Warrington, E., Sanders, M. & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, 97, 709-728.

Wilkes, K. (1988). *Real people: Personal identity without thought experiments*. New York: Oxford University Press.

Williams, M. & Mattingley, J. (2003). Unconscious perception of non-threatening facial emotion in parietal extinction. *Experimental Brain Research*, 154(4), 403-406.

Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wilson, T., Lindsey, S. & Schooler, T. (2000). A model of dual attitudes. *Psychological Review*, 107, 101-26.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103-128.

Winter, L. & Uleman, J. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, 47, 237-252.

CURRICULUM VITAE

Education

Dates	*School*	*Degree*
2004-2011	Rutgers University	Ph.D., Philosophy Graduate Certificate, Cognitive Science
1998-2002	Kenyon College	B.A., Philosophy and Political Science

Positions

Dates	*Position*
2010-2011	Mellon Dissertation Fellow
2009-2010	Teaching Assistant
2005-2009	Jacob K. Javits Fellow
2004-2004	Andrew W. Mellon Fellow

Publications

Goldman, A. & Shanton, K. (Forthcoming). The Case for Simulation Theory. In *Handbook of Theory of Mind* (Eds.), A. Leslie & T. German.

Shanton, K. (2011). Memory, Knowledge and Epistemic Competence. *Review of Philosophy and Psychology*, 2(1), 89-104.

Shanton, K. & Goldman, A. (2010). Simulation Theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4), 527-38.