FIVE

Consciousness and Free Will (II): Transparency, Infallibility, and the Higher-Order Thought Theory

Despite the arguments in the previous chapter there are still many who believe consciousness not only provides us with evidence that we are free but that consciousness itself is the vehicle by which freedom is secured. We typically feel that our conscious selves are in the driver's seat and that we're able to exercise agent-causal control over our actions through our conscious intentions, decisions, and volitions. Upon introspection, we also feel as though we are not causally determined to act as we do. Whereas compatibilists often avoid discussions of consciousness, libertarians generally rest their entire case on the phenomenology of conscious agency and our feeling of freedom. It is not uncommon to appeal to conscious experience, especially the supposed consciousness we have of our own freedom, as evidence for the reality of free will. As Simon Blackburn points out, "Consciousness of freedom seems closely allied to any kind of consciousness at all" (1999, 82). We all seem to be aware of our own freedom in the very act of deliberation, choice, and action. This, I take it, is what gives strength to the libertarian argument and what makes it difficult, perhaps impossible, for us to admit we are not free. The way we experience our own minds, and the fact that we do not experience the multitude of unconscious determinants acting on us, gives us the false sense that the conscious self is in control.

Since our sense of free will is primarily a libertarian one, I would here like to explore the libertarian approach to consciousness and mentality as an entry point into my account of the illusion of free will. I will lay the groundwork for my account of the illusion in this chapter and then build on it in subsequent chapters. In this chapter, I attempt to explain one

feature of consciousness—the apparent transparency and infallibility of consciousness—that contributes to the cognitive illusion of free will. In chapters 6 and 7, I will explore additional aspects of consciousness that further contribute to the illusion—e.g., the apparent spontaneity of intentional states, the feeling that we consciously cause and initiate behavior, and the nature of self-consciousness. Although the scientific data reveals a pervasive adaptive unconscious that controls a good deal of our day-to-day lives, our own experience supports a different view of the mind. Phenomenologically we feel as though we, our conscious selves, are in complete control of our intentional/voluntary behavior and that we are in the best position to judge why we act as we do. I will argue that this is due, in part, to a belief in the transparency and infallibility of consciousness. I argue that our belief in the introspective transparency and infallibility of consciousness (which is supported by phenomenology and the way we experience our own mental states), coupled with a failure to introspect any deterministic processes underlying our own decision making, leads us to (wrongly) infer that we are free and causally undetermined.

In this chapter, I also examine two different approaches to consciousness. I first examine theories of consciousness that deny the existence of unconscious mental states. Although these theories are empirically unwarranted given the psychological data presented in the last chapter, they are worth analyzing because they are supported by phenomenology and are more accommodating to free will. Analyzing this approach to consciousness can help us understand why the belief in free will is so powerful. As an alternative, I introduce and defend the *higher-order thought* (or HOT) theory of consciousness as developed by David Rosenthal (2005a). I argue that this theory has several virtues, not the least of which is that it can explain why certain mental states are conscious and not others. I will use the HOT theory in this chapter to explain away the *apparent* transparency and infallibility of consciousness, and I will use it in subsequent chapters to analyze additional aspects of the illusion of free will. The HOT theory is one of the few authentic theories of consciousness available in the literature. Although consciousness has become a hot topic over the last few decades, there are surprisingly few theories out there that venture to formulate necessary and sufficient conditions for consciousness. The HOT theory is one of those theories and it is the one I wish to defend here.

In addition to presenting and defending the HOT theory, I will also consider a larger question: namely, what causal powers does consciousness actually add? Since I have already argued that higher mental processes are largely controlled and determined by unconscious and automatic processes, this will be an important question. I will argue that although conscious mental states do possess distinctive causal powers, these powers are more limited than we typically think and are no differ-

ent in *type*. I maintain that although consciousness brings with it a *feeling of freedom* it does not bring with it actual freedom. And I conclude that our *subjective feeling of freedom* is nothing but a conscious illusion.

5.1 CONSCIOUSNESS AND FREEDOM: THE INTROSPECTIVE ARGUMENT FOR FREE WILL

In the free will debate, libertarians put a great deal of emphasis on our *conscious feeling of freedom* and our introspective abilities. In fact, many libertarians have suggested that our introspection of the decision-making process, along with our strong feeling of freedom, provides some kind of *evidence* for the existence of free will. As Ledger Wood describes the libertarian argument: "Most advocates of the free will doctrine believe that the mind is directly aware of its freedom in the very act of making a decision, and thus that freedom is an immediate datum of our introspective awareness. 'I feel myself free, *therefore*, I am free,' runs the simplest and perhaps the most compelling of the arguments for freedom" (1941, 387). We can call this the *introspective argument* for free will. The introspective argument essentially maintains that, upon introspection, we do not *seem* to be causally determined—instead, we *feel* that our actions and decisions are freely decided by us—hence, we *must* be free. Libertarians, especially agent-causal theorists, take this introspective datum as their main evidence in support of free will. Timothy O'Connor, for example, writes:

> [T]he agency theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favor doing so; it seems to be the case, rather, that *I* produce my decision *in view of* those reasons, and could have, in an unconditional sense, decided differently . . . Just as the non-Humean is apt to maintain that we not only perceive, e.g., the movement of the axe along with the separation of the wood, but the axe *splitting* the wood . . . , so I have the apparent perception of my actively and freely deciding to take Seneca Street to my destination and not Buffalo instead. (1995b, 196)

Richard Taylor, another leading agent-causal theorist, maintains that there are two introspective items of data: (1) That I *feel* that my behavior is sometimes the outcome of my deliberations, and (2) that in these and other cases, I *feel* that it is sometimes up to me what I do (1992, ch.5). He then concludes: "The only conception of action that accords with our data is one according to which people—and perhaps some other things too—are sometimes, but of course not always, self-determining beings; that is, beings that are sometimes the cause of their own behavior" (1992, 51). C.A. Campbell makes a similar point with regard to moral deliberation:

> The appeal is throughout *to one's own experience* in the actual taking of the moral decision as a *creative* activity in the situation of moral temptation. "Is it possible," we must ask, "for anyone so circumstanced to *dis*believe that he could be deciding otherwise?" The answer is surely not in doubt. When we decide to exert moral effort to resist temptation, we feel quite certain that we *could* withhold the effort; just as, if we decided to withhold the effort and yield to our desires, we feel quite certain that we *could* exert it—otherwise we should not blame ourselves afterwards for having succumbed. (1957, 169)

The *introspective argument* is therefore important because "all libertarians assign introspective evidence some role, for it is our feeling of metaphysically open branching paths that is the raison d'être of libertarian freedom" (Ross 2006, 135).[1]

This kind of argument only works, however, if we assume the data is veridical. But how do we know that our feeling of freedom isn't an illusion? How do we know that what we introspect is accurate? Such arguments, I maintain, fail to *prove* that our phenomenological appearances accurately represent reality. For one, it is a mistake to think that one could establish a metaphysical conclusion from phenomenology alone. More than an appeal to our introspective experience is needed to prove that we actually *enjoy* agent-causation. Secondly, and perhaps more importantly, I will now argue that there is reason to doubt the reliability of such introspective evidence. Despite such concerns, however, libertarians seldom question the "I feel myself free, therefore, I am free" argument. In fact, the introspective argument can be found throughout the literature. Although such arguments may not prove we are actually free, they do reveal that the libertarian conception of agency—a conception I have argued is shared by common sense—is deeply rooted in our conscious feeling of freedom and a belief in the accuracy of introspection. Given this, it is important that we investigate both the role of consciousness and the accuracy of introspection. I will argue that a closer examination of these issues will reveal that the nature of consciousness, rather than supporting free will, further impugns it.[2]

Libertarianism and Consciousness

Although libertarians put a great deal of emphasis on consciousness when it comes to introspecting our own freedom, they ironically overlook the importance of consciousness when it comes to explaining its role in *producing* free actions. O'Connor, for example, seems to be aware of this shortcoming when he writes:

> Something the philosopher ought to be able to provide some general light on is how consciousness figures into the equation. It is a remarkable feature of most accounts of free will that they give no essential role to conscious awareness. One has the impression that an intelligent

> automata could conceivably satisfy the conditions set by these accounts—something that is very counterintuitive. (2000, 122)

I share O'Connor's surprise at the fact that consciousness has not played a larger role in accounts of free will, especially given the obvious importance of conscious awareness. It truly is counterintuitive to think that one could exercise free will unconsciously. As I argued in the previous chapter, any successful account of free will must explain the role consciousness plays in the exercise of free agency and account for the prominence we give to *conscious will*. According to folk psychology (as well as many philosophers and psychologists) it is logically inconceivable to imagine an *automaton*—a creature that lacks all conscious awareness—that has freedom. An intelligent automaton *cannot* and *should not* be the paradigm of a free agent, for the picture one gets of an automaton is that of an intelligent robot; a robot that perhaps can learn to adapt to its environment, a robot that may even have a certain amount of flexibility, but one that completely lacks freedom. When we say that an action is free, we typically mean (among other things) that it was the result of a voluntarily choice, *consciously willed*. Conscious will is believed to be an essential aspect of free will.

Given that consciousness seems to be a necessary condition for freedom, why have so many accounts of free will overlooked it? O'Connor speculates at an answer: "That accounts of free will fail to provide an essential role for consciousness is nonetheless not surprising, given that its basic biological functions are presently quite mysterious to most theorists" (2000, 122). Although I agree that not *all* of the biological functions of consciousness are presently known, it is a mistake to think that none of them are. Consciousness research has a long way to go, but at a minimum we know that consciousness plays an important role in monitoring our internal and external environments, contemplating long-range action plans, and in facilitating memory formation and reasoning.[3] In addition, libertarians and compatibilists cannot simply neglect the importance of consciousness because the whole story is not yet in. Libertarians especially have to take the active role of showing that consciousness somehow imparts to agents a power not possessed by automata or unconscious creatures.

Immediately after the previous quote, O'Connor continues: "Another aspect of the puzzle is that whereas various suggestions have been put forth concerning what specific function or functions consciousness serves, it is readily imaginable that many of these functions can be carried out by automata" (2000, 122). If libertarian accounts of freedom are to be successful, they *must* show that this is not the case. That is, they need to show that *one* of the functions of consciousness is that it somehow exercises or facilitates free will. To his credit, O'Connor recognizes this point. He states:

> It is highly plausible that this self-determining capacity strictly requires conscious awareness. This appears to follow from the very way in which active power has been characterized as structured by motivating reasons and as allowing the free formation of executive states of intention in accordance with one of the possible courses of action represented to oneself. (I am tempted to think that one should be able to explicitly demonstrate the absurdity of supposing an agent-causal capacity as being exercised entirely unconsciously). (2000, 122)

Given the requirement of conscious awareness, then, it is a sad state of affairs when libertarians, like O'Connor and others, dedicate no more than a few lines to the issue. O'Connor himself only presents one, very vague proposal. He claims, "The agency theorist can conjecture that *a* function of biological consciousness, in its specifically human (and probably certain other mammalian) manifestations, is to subserve the very agent-causal capacity I sketched in previous chapters" (2000, 122). Beyond this, O'Connor does not explain *how* or *in what way* consciousness 'subserves' these presumed agent-causal powers. And this general failure can be found throughout the libertarian literature. Essentially libertarians give us a mere promissory note for the key component of their theory. They fail to provide us with any substantial account of how consciousness carries out this key biological function—i.e., the agent-causal exercise of libertarian freedom.

It seems then that libertarians lack a complete story. On the one hand they appeal to our conscious feeling of freedom as evidence of free will, while on the other hand they neglect to explain the role and importance of consciousness. O'Connor's comments simply amount to the following two claims: (1) That the "self-determining capacity [required for libertarian freedom] strictly requires conscious awareness"; and (2) *somehow* consciousness aids in this capacity. This exposes, I believe, another major problem with libertarianism (and, in general, most defenses of free will)—they typically fail to explain the role of consciousness in the exercise of free will. (In some ways compatibilists are even worse since they usually avoid discussions of consciousness altogether.) Since it is not my job to speculate on how consciousness can aid in self-determination—in fact, I will explicitly attempt to show that it cannot—I will instead focus on the nature of conscious awareness to see what else it can tell us about free will. I will attempt to show that certain features of consciousness lead us to impute more control to the conscious self, and put more faith in the introspective argument, than we should.

Let us return to the introspective argument for a moment. As I have already stressed, from the fact that I *feel* free, it does not necessarily follow that I *am* free. The feeling could be an illusion. What this argument does show, however, is that people often infer their own freedom from their introspective phenomenology of freedom. Why is this so? I propose

that people implicitly believe that they have access to all the causal factors and processes underlying their own decision-making. If people were to believe in such introspective transparency, then it would be appropriate, given the above phenomenology, for them to infer that they are undetermined. For if one introspects no deterministic processes underlying one's decision making, and one also thinks that if there *were* a deterministic process one *would* introspect it, one would infer that there is no deterministic process.[4] I therefore argue that a standing phenomenological belief in the introspective transparency and infallibility of consciousness, coupled with a failure to introspect any deterministic processes underlying our own decision making, helps contribute to our sense of free will.[5] From the first-person point of view, we feel as though consciousness is immediate, direct, transparent, and infallible. The *apparent* immediacy, transparency, and infallibility of consciousness leads us to assume a kind of first-person authority where we believe that there can be no mental causes for our actions other than the ones we are aware of. Because we do not experience the multitude of unconscious determinants at work, and because we (wrongly) believe that we *would* be aware of such determinants if they were present, we conclude that no such determinants exist. The phenomenology tricks us here into thinking that it is our conscious will alone that is in control.[6]

This proposal assumes, of course, that consciousness is *not* transparent and does *not* provide us with infallible knowledge of our own mental processes. Am I right to assume this? Clearly, from a first-person point of view, we *feel* as though we are immediately and infallibly connected to our own minds. Can we be wrong about this? Many find it difficult, even impossible, to question our introspective authority. This, of course, is not surprising given the nature of conscious experience. It is important, however, to further examine the nature of consciousness and mentality to see whether we really are transparently and infallibly aware of the inner workings of our mind; for it is our belief in the transparency and infallibility of consciousness that gives the introspective argument whatever power it possesses and contributes to our sense of freedom.

5.2 TWO CONCEPTS OF CONSCIOUSNESS

I would now like to examine two different approaches to consciousness; one that supports the reasoning behind the introspective argument, and one that questions its core assumptions.[7] Given that the majority of support for free will comes from our introspective awareness of the decision-making process, and consciousness appears to be a necessary condition for free will, it would seem that an account of the mind which claims that all mental states are conscious states would be more accommodating to defenders of free will. Such an account of the mind can be traced back to

René Descartes. One can find at the heart of Descartes' philosophy of mind three main theses:

1. That the mind and body are two mutually exclusive, interacting substances—the mind being completely nonphysical.
2. That there is nothing in our mind of which we are not conscious; i.e., all mental states are conscious states; and
3. Our knowledge of our own mental states is certain and infallible; our judgments about them cannot be erroneous.

These three theses comprise the core of Descartes' philosophy of mind. I have already presented an argument against the first of these theses. I have argued, following Kim and Papineau, that worries over mental causation show that interactive substance dualism is an indefensible position. In fact, such worries have caused most to give up the thesis—Antonio Damasio has even dubbed it "Descartes' error" (1994). I would now like to focus on the latter two theses.

Despite a retreat from the metaphysics of substance dualism, the rest of the Cartesian concept of mind remains largely intact when it comes to theorizing about consciousness and free will. Theses (2) and (3) combined amount to the claim that all mental states are conscious and that such consciousness is infallible. Essentially this is the belief in the *transparency* and *infallibility* of consciousness.[8] From a first person point of view these two theses seem compelling. We are conscious of our mental states in a way that seems, at least subjectively, to be direct, immediate, and infallible.

What we need to investigate is whether our phenomenology, which seems to support these two theses, is accurate. I will argue that it is not. From a first-person point of view, it may *seem* as though we are aware of all our mental states and processes—including reasoning and decision making—in an immediate, direct, and infallible way, but from a third-person point of view we can often see that this is not the case. I believe that it is partly because consciousness appears transparent from the first-person point of view that we impart so much power to the conscious will. The fact that, subjectively, mental functioning *appears* transparent to consciousness leads us to attribute more power to consciousness than it actually has. The way we are connected to our own minds produces a misleading feeling of confidence. If consciousness is *not* transparent, then the introspective argument for free will lose its force.

In this section, I will argue that a Cartesian conception of consciousness is neither theoretically desirable nor empirically supported. I will then turn, in the following section, to an alternative conception of consciousness: the *higher-order thought theory* of consciousness. I will argue that consciousness is best viewed as extrinsic to mental states, and that what makes a mental state conscious is one's being *conscious of* that state in some suitable way. I will end the chapter with a look at what this

alternative conception of consciousness tells us about free will and some speculation on the function of consciousness.

Let us begin with the claim that all mental states are conscious states. This belief is a main tenet of the Cartesian concept of mind. Descartes famously writes that:

> [T]here can be nothing in the mind, in so far as it is a thinking thing, of which it is not aware, this seems to me to be self-evident. For there is nothing that we can understand to be in the mind, regarded in this way, that is not a thought or dependent on a thought. If it were not a thought or dependent on a thought it would not belong to the mind *qua* thinking thing; and we cannot have any thought of which we are not aware at the very moment when it is in us. (CSM, II:171)[9]

And in the *Second Set of Replies*, Descartes defines "thought" as follows: "I use this term to include everything that is within us in such a way that we are immediately aware of it" (CSM, II:113; AT, VII:160).[10] Since the reference here to thoughts was meant to cover all mental states of whatever kind, including intentional states and sensory states, these remarks are representative of the Cartesian idea that all mental states must be conscious states.[11]

This conception of mentality and consciousness has influenced many philosophers. As Rosenthal points out, "This view is epitomized in the dictum, put forth by theorists as otherwise divergent as Thomas Nagel (1974: 174) and Daniel Dennett (1991: 132), that the appearance and reality of mental states coincide" (2004b, 17; see also Caruso 2005). Not only does it claim that consciousness is an essential property of mental states, but also that consciousness is the mark of the mental. For on the Cartesian concept of mind, what makes a state a mental state is its being a conscious state. States that are not conscious are also not mental. This, however, has significant theoretical drawbacks. If consciousness is what makes a state a mental state, consciousness will not only be an intrinsic, nonrelational property of all mental states, it will also be unanalyzable. Rosenthal, for example, has argued that if being mental means being conscious, we can invoke no mental phenomenon whatever to explain what it is for a state to be a conscious state. And "Since no nonmental phenomenon can help, it seems plain that, on the Cartesian concept of mentality, no informative explanation is possible of what it is for a mental state to be conscious" (1986, 31). Since the Cartesian concept of mind tacitly conflates mentality and consciousness—thereby making consciousness essential to all mental states—no reductive explanation of consciousness can be given in terms of other higher-level cognitive or mental processes. And this precludes giving any informative, nontrivial account of what such consciousness consists in (see Rosenthal 1986, 2002c).

The main difficulty with equating mind and consciousness has to do with understanding the nature of consciousness. If mental states are all conscious, argues Rosenthal, we will simply be unable to understand the very nature of consciousness itself. The problem is the following:

> Suppose mental states are all conscious. How could we then explain what it is for a mental state to be conscious? There are two ways we might proceed. One of these appeals to something mental; we explain what it is for one mental state to be conscious in terms of other mental states or processes. This will not do. If all mental states are conscious states, then the other mental phenomena to which our explanation appeals will themselves be conscious. So this kind of explanation results in a vicious regress. We cannot explain what it is for any mental state to be conscious except in terms of another mental state, whose being conscious itself requires explanation. (2002c, 235)

Since this is unacceptable, the only alternative is to explain what it is for a mental state to be conscious without appealing to any other mental phenomena. The problem with this, however, is that it is highly unlikely that we can understand what it is for a mental state to be conscious appealing only to things that are themselves not even mental. As Rosenthal argues:

> Consciousness is the most sophisticated mental phenomenon there is and the most difficult to understand; nothing in nonmental reality seems to be at all suited to help us grasp its nature. If we are to have any informative explanation of what it is for a mental state to be conscious, it is all but certain that it will have to make reference to mental phenomena of some sort or other. (2002c, 236)

It is important to realize that what we are after here is not a scientific explanation. What we want, instead, is "to understand just what the phenomenon is that a scientific theory might then explain" (Rosenthal 2002c, 235). Since we are looking for a theoretical account of what makes a mental state a conscious state, and not a scientific account, I agree with Rosenthal that an appeal to other mental phenomenon is necessary.[12]

It would seem then that if all mental states are conscious, we can give no informative account of what such consciousness consists in—i.e., we would be unable to explain what makes a mental state conscious. This, I maintain, is a serious problem for any theory of consciousness that equates mind and consciousness. But what is equally troubling, or perhaps even more troubling, is what accepting this equivalence means for understanding the mind itself, not only consciousness. If we were to equate mind and consciousness, we would then have to understand mental processes in terms of consciousness. But doing so would also prevent us from ever developing an informative account of mind (see Rosenthal 2002c, 237). We would be unable to investigate mental processes without at the same time investigating conscious processes. This, I believe, is not

only theoretically unacceptable, but given everything I argued in the previous chapter, it is also empirically unjustifiable.

There is more than ample reason, as we've seen, to believe that not *all* mental states are conscious states. Many types of mental states—such as thoughts, desires, beliefs, judgments, goals, and intentions—often occur without being conscious. Both common sense and cognitive science typically posits mental states that are not conscious to explain certain behaviors and cognitive capacities. The most widely accepted of these unconscious mental states are intentional states. There are not only experimental results which provide good reason to hold that beliefs and desires exist that are not conscious, but everyday folk psychology makes much use of intentional states that are not conscious to explain the actions of others. In fact, the majority of philosophers, *pace* John Searle (1990, 1992), now agree that there are nonconscious intentional states. One could even argue that the majority of our intentional states are probably nonconscious.

As I noted earlier, the work of Timothy Wilson, John Bargh, Benjamin Libet, and others have shown that the higher mental processes that have traditionally served as quintessential examples of choice and free will—such as goal pursuits, judgment, interpersonal behavior, and action initiation—can and often do occur in the absence of conscious choice or guidance. It is no longer believed that *only* lower-level processing—or what we can call sub-mental processing (such as perceptual processing)—can occur outside the reach of consciousness. There is now growing evidence that a great deal of higher-level mental functioning is also nonconscious. Psychologists and cognitive (and social-cognitive) scientists have accumulated a great deal of evidence for determinism by demonstrating that high-level mental and behavioral processes can proceed without the intervention of conscious deliberation and choice. All of this mounting research, I believe, proves that high level unconscious cognitive states—states that are best described as *mental*—actively and frequently play a role in human behavior. Some of this research was discussed in chapter 3.

I should point out that when talking about nonconscious mental states, I do not mean simply to be talking about dispositional states; states that are disposed to be occurrent conscious states. I mean to be making the stronger claim that these are occurrent nonconscious states—states that influence behavior and interact with other mental states, both conscious and nonconscious. I also believe discussion of unconscious mental states should not be limited simply to intentional states. Although nonconscious intentional states are more widely acknowledged, there is, indeed, good reason to believe that unconscious sensory states also exist (see Rosenthal 1986, 1991a, 1997; Grahek 2007; Caruso 2005). Following Rosenthal, I believe:

> [T]here is reason to hold that the sensory qualities characteristic of our conscious sensations occur even when sensations of the relevant types fail to be conscious, as in peripheral vision and subliminal perceptions and in laboratory contexts such as experiments involving masked priming. When sensations occur without being conscious, they often still affect our behavior and mental processes in ways that parallel the effects of conscious sensations. (2002c, 242)[13]

We have already seen, for example, how unconscious perception can affect behavior in subliminal perception and masked priming experiments. These would be examples of unconscious sensory states. Many theorists also posit nonconscious sensory states—states with sensory qualities—to explain cases of so-called "blindsight" (see Weiskrantz 1986, 1997; Caruso 2005).

The Cartesian thesis, then, that all mental states are conscious states—though perhaps supported by phenomenology—is simply false. Timothy Wilson has even compared it to "Descartes' error" of Cartesian dualism. He writes:

> Descartes made a related error that is less well known but no less egregious. Not only did he endow the mind with a special status that was unrelated to physical laws; he also restricted the mind to consciousness. The mind consists of all that people consciously think, he argued, and nothing else. This equation of thinking and consciousness eliminates, with one swift stroke, any possibility of nonconscious thought—a move that was called the "Cartesian catastrophe" by Arthur Koestler and "one of the fundamental blunders made by the human mind" by Lancelot Whyte. Koestler rightly notes that this idea led to "an impoverishment of psychology which it took three centuries to remedy." (2002, 9-10)

Theories of consciousness which still maintain that all mental states are conscious—like those of Searle (1990, 1992), Dretske (1995), and Tye (1995)—therefore remain a stumbling block in the way of progress. What we need is a theory of consciousness that is able to explain why certain mental states are conscious and not others. As Rocco Gennaro puts it, "One question that should be answered by any viable theory of consciousness is: What makes a mental state into a conscious one?" (2004, 1). Any theory that is unable to answer this fundamental question leaves, what Rosenthal has called, *state* consciousness completely unexplained.[14]

Knowing Thy Self: Consciousness and Self-Reports

In addition to the assumption that all mental states are conscious, libertarian and folk-psychological accounts of consciousness usually make the related assumption that consciousness provides us with infallible knowledge of our own minds. The claim of infallibility is another part of the traditional Cartesian concept of mind.[15] From a first-person

point of view, this assumption seems to makes sense. Who else, we feel, is in a better position to know which mental states we are in than ourselves? It is often assumed, almost at a definitional level, that we are immediately, directly, and infallibly connected to the content of our own minds. From a first-person point of view, it never seems as though consciousness and mentality come apart. Subjectively, it never seems to us that consciousness mischaracterizes or misidentifies the mental states we are in. It is hard for us to believe that our consciousness can mislead us about the nature of our own minds or that we can be in mental states that we are unaware of.

Although this is undoubtedly how things seem from a first-person point of view, I do not think we can rely on phenomenology alone to the exclusion of all other information. There is a great deal of research suggesting that we are not always the best judges of what's going on in our own minds. As Rosenthal points out:

> [C]onsciousness does not always represent our mental states accurately. Consciousness seems infallible because it never shows itself to be mistaken and it's tempting to think that there's no other way to know what mental states one is in. But consciousness is not the only way to determine what mental state one is in, and there is sometimes compelling independent evidence that goes against what consciousness tell us. (2004b, 27)

Researchers, for example, are increasingly realizing that the mental states and processes that they are interested in measuring are not always consciously accessible to their participants, forcing them to rely on alternative methods (see Wilson 2003). The introspective method—i.e., the method of relying on the introspective reports of subjects—has, in fact, come under attack numerous times throughout the history of psychology (e.g., Nisbett and Wilson 1977; Lieberman 1979; Jack and Roepstorff 2002). In attitudes research, for example, a number of researchers now argue that people can simultaneously possess different implicit and explicit attitudes toward the same object, with self-reports measuring only the explicit attitude (e.g., Wilson, Lindsey, and Schooler 2000). This has led some to develop implicit measures to explore the nature of these attitudes and people's awareness of them (Greenwald, McGhee, and Schwartz 1998).[16]

Our ability to know our own mental states is limited and fallible. People have access to many of their mental states, no doubt, but there is also a pervasive adaptive unconscious that is often inaccessible via introspection. In addition, consciousness, which is accessible to introspective reports, does not always represent our mental states and processes accurately. Individuals often confabulate stories for why they do certain things. When this happens, one's first-person reports fail to match the actual causes for their action. This has been shown to happen, for exam-

ple, in hypnotized subjects. After being hypnotized, subjects can enact a posthypnotic suggestion—e.g., "when you awake you will immediately crawl around on your hands and knees." When asked what they are doing, subjects almost immediately generate a rationale—"I think I lost an earring down here" (Gazzaniga 1985; Hilgard 1965; Estabrooks 1943). From a first-person point of view, these individuals are conscious of a particular reason for why they are doing what they are doing, but from a third-person point of view we can see that this is not the real cause of their action. Similar examples of confabulation have also been found in "split brain" patients (Gazzaniga and LeDoux 1978) and patients with Korsakoff's syndrome—a form of organic amnesia where people lose their ability to form memories of new experiences (Sacks 1987).[17]

Although it may be tempting to think such confabulation is limited to these rare occasions, some theorists have suggested that similar confabulation occurs throughout everyday life (see Nisbett and Wilson 1977; Gazzaniga and LeDoux 1978; Wilson 2002). These theorists argue that our conscious selves often do not fully know why we do what we do and thus have to confabulate stories and create explanations. In one of the most famous papers on the subject, Nisbett and Wilson (1977) placed subjects in identical situations save for the fact that one or two key features were varied. They observed that although these key features influenced people's judgments or behavior, when asked to explain why they responded the way they did, subjects remained unaware of the varied features and instead confabulated different explanations for their behavior.

In one study, for example, Nisbett and Wilson attempted to see if people could express accurately all the reasons why they preferred one pair of panty hose to another. In the study, conducted in a commercial establishment under the guise of a consumer survey, passersby were invited to evaluate four identical pairs of nylon stockings. The panty hose were arranged neatly on a table labeled A, B, C, and D, from left to right. Nisbett and Wilson found a pronounced left-to-right position effect, such that the right-most object in the array was heavily over-chosen by a factor of almost four to one. They knew that this was a position effect and not that pair D had superior characteristics because all the pairs of panty hose were identical—a fact that went unnoticed by almost all the participants. When asked about the reasons for their choice, no subject ever mentioned spontaneously the position of the article in the array. And, when asked directly about a possible effect of the position of the article, virtually all subjects denied it. Instead of accurately reporting why they chose their preferred pair, people confabulated reasons having to do with superior knit, sheerness, or elasticity. Studies like this seriously question the accuracy of consciousness awareness, because they reveal that consciousness does not provide us with transparent and infallible knowledge of our

own minds. We may consciously think we are doing something because of reason X, when in reality we are doing it because of reason Y.[18]

What does this mean for the introspective argument for free will? I believe it shows that we cannot rely on our conscious experience *alone* to determine the causes of our actions. We are often unaware of important causal determinants. The fact that we do not *feel* causally determined, or that we are not consciously aware of the various internal and external influences on our behavior, does not mean such determinants do not exist. Worse still, if consciousness can confabulate and/or misrepresent the causes for our choices and/or actions, then to rely on such conscious data to infer our own freedom would be a mistake. Whatever persuasiveness the introspective argument originally had depended on the assumption that we had direct, infallible access to our own decision-making process. The argument assumes that consciousness reveals everything about our mental functioning, or at least everything relevant to the issue at hand. This, however, is not the case. What we are conscious of, and hence what we can report on, is not always in line with what is otherwise going on mentally.

We have now seen that identifying mind and consciousness not only makes it impossible to give an informative account of what consciousness consists in, it is also incompatible with everything we know about mental processes. Recent research into automaticity and the adaptive unconscious has revealed that sophisticated, higher mental processes can occur without conscious awareness, control, or intervention. We have also seen that consciousness can at times confabulate stories and misidentify the states we are in. What we need, then, is a conception of consciousness that does justice to these two insights. We need an account of consciousness which explains why consciousness *appears* transparent from a first-person point of view, yet also explains why it is not. As Rosenthal puts it, "Consciousness does reveal the phenomenological data that a theory of consciousness must do justice to. But to save these phenomena, we need only explain why things appear to consciousness as they do; we need not also suppose that these appearances are always accurate" (2004b, 31). I will now introduce a theory of consciousness which explains state consciousness in terms of higher-order awareness. I will argue that consciousness is best viewed as extrinsic to mental states, and that what makes a mental state conscious is one's being conscious of that state in some suitable way. This account of consciousness, while accurately capturing the phenomenology, will explain how there can be unconscious mental states. It will also explain how we can misrepresent or even on occasion confabulate the mental states we are conscious of.

5.3 THE HIGHER-ORDER THOUGHT (HOT) THEORY OF CONSCIOUSNESS

Let me begin by sketching in its most basic outline the *Higher-Order Thought* (HOT) theory of consciousness. David Rosenthal has made the clearest and best case for the HOT hypothesis in a series of cogent and convincing papers (1986, 1991a, 1993a, 1993c, 1993d, 1997, 2002c, 2002d, 2003, 2004b). Accordingly, I will concentrate on Rosenthal's version of the theory, focusing specifically on how it deals with the research on automaticity and the adaptive unconscious. I will argue that the HOT theory is particularly well suited to account for how there can be unconscious mental states, as well as to accommodate disparities between our higher-level mental processes and our first-person reports of those processes.

The HOT theory belongs to a larger class of theories known as *higher-order* (or HO) theories of consciousness. Such theories can be traced as far back as John Locke (1689). Recently, HO theories have been presented by a number of philosophers. Besides Rosenthal, HO theories have been advanced by Armstrong (1968, 1981), Lycan (1996), Carruthers (1996, 2000), and Gennaro (1996, 2005). Gennaro describes the basic idea behind HO theories as follows: "In general, the idea is that what makes a mental state conscious is that it is the object of some kind of higher-order representation (HOR). A mental state M becomes conscious when there is a HOR of M. A HOR is a 'meta-psychological' state, i.e., a mental state directed at another mental state" (2004, 1). According to HO theories, then, my desire to get this chapter done becomes conscious when I somehow become *aware* of that desire; i.e., when I have a HOR of that desire. HO theories have intuitive appeal since a state of which one is in no way aware does not intuitively count as conscious. As Rosenthal writes:

> If an individual is in a mental state but is in no way whatever conscious of that state, we would not intuitively count it as a conscious state. So a state's being conscious consists of one's being conscious of it in some suitable way. . . . It is this equivalence of a state's being conscious with one's being conscious of it in some suitable way that points toward a higher-order theory of what it is for a mental state to be conscious. (2004b, 17)

According to HO theories, we can explain a state's being conscious in terms of a higher-order state's being directed on that state. This hierarchical or iterative structure is what distinguishes HO theories from other theories of consciousness—especially first-order representational (FOR) theories, like those of Tye (1995) and Dretske (1995).[19]

There are essentially two types of higher-order theories, differing on how they understand the HOR. There are higher-order *thought* (HOT) theories—like those of Rosenthal (2005a), Carruthers (1996, 2000), and Gennaro (1996)—and there are higher-order *perception* (HOP) or higher-

order *sensory* theories—like those of Lycan (1996) and Armstrong (1968, 1981). These two HO theories differ over the nature of the higher-order representation. HOT theorists, like Rosenthal, argue that the higher-order state should be viewed as a thought. HOP theorists, on the other hand, argue that the HOR is closer to a *perceptual* or *experiential* state of some kind. The central difference between these two views is over the need for conceptual content. More specifically, HOT theorists maintain that the HOR is a *cognitive* state involving some kind of conceptual component, whereas higher-order perception or sensory theories argue that the HOR is closer to a *perceptual* or *experiential* state of some kind which does not require the kind of conceptual content invoked by HOT theorists. Because of this difference, and largely due to Kant (1781), the latter are sometimes referred to as "inner sense" theories as a way of emphasizing this sensory or perceptual aspect. So whereas the HOT theory contends that a mental state is conscious just in case it is the object of a higher-order thought (i.e., a cognitive state with conceptual content), the HOP (or "inner sense") theory holds that the HOR is a kind of internal scanning or monitoring by a quasi-perceptual faculty.

Some philosophers have argued that the difference between these theories is perhaps not as important or as clear as some think it is (Gennaro 1996; Van Gulick 2000). Others, like Guzeldere (1995), have argued that the HOP theory ultimately reduced to the HOT theory. I, myself, believe that the HOT theory, as developed by Rosenthal, has distinct advantages over the HOP theory. I also maintain that Rosenthal's version of the HOT theory is superior to those of Carruthers and Gennaro. Some of my reasons for believing this will come out below, but for the most part I will not discuss these other HO theories. I will instead focus on Rosenthal's version of the HOT theory, pointing out, where possible, how it differs from these other theories. For a more detailed account of why Rosenthal's version of the HOT theory is superior to these other HO accounts, see Rosenthal's "Varieties of Higher-Order Theory" (2004b).

Although there are significant differences between these accounts, all HO theories agree that what makes a mental state conscious is its relation to another, higher-order, mental state. Because of their hierarchical nature, "HO theories are also attractive to some philosophically inclined psychologists and neuroscientists partly because they suggest a very natural realization in the brain structure of humans and other animals" (Gennaro 2004, 2). See, for example, Rolls (1999), Weiskrantz (1997), and Lau (2007, 2010). Gennaro describes the basic appeal of HO theories as follows: "At the risk of oversimplification, if we think of the brain as developing layers upon layers corresponding to increasing sophistication in mental ability, then the idea is that mental states corresponding to various 'higher' areas of the brain (e.g., cortex) are directed at various 'lower' states rendering them conscious" (2004, 2). In fact, a number of HO theorists have maintained that first-order perceptual representations,

for example, depend on neural activity in early sensory regions, whereas higher-order representations depend on neural activity mainly in prefrontal (and parietal) cortex (e.g., Lau and Rosenthal 2011; Lau 2010; Kriegel 2009). And although not a necessary condition of such theories, this empirical interpretation of HO theories has recently received support from emerging findings in cognitive neuroscience, giving the view substantial empirical credibility and an advantage over its competitors (see Lau and Rosenthal 2011). It's important to point out, however, that HO theories themselves do not, in general terms, attempt to reduce consciousness *directly* to neurophysiological states. As Gennaro describes:

> Unlike some other theories of consciousness (Crick & Koch, 1990; Crick, 1994), they are not reductionist in the sense that they attempt to explain consciousness directly in physicalistic (e.g., neurophysiological) terms. Instead, HO theories attempt to explain consciousness in *mentalistic* terms, that is, by reference to such notions as 'thoughts' and 'awareness.' . . . HO theorists are normally of the belief that such mental states are identical with brain states, but they tend to treat this matter as a further second step reduction for empirical science. (2004, 2)

HO theories, then, provide a mentalistic reduction of consciousness. Although this is different from a physicalistic reduction, HO theories are more accommodating to materialist accounts of the mind than are Cartesian accounts.

A mentalistic reduction of consciousness would be the first step in reducing consciousness to brain states. Armstrong, for example, endorses a HO theory of consciousness for the mentalistic reduction, and then famously proposes a *causal-functionalist* theory (1966, 1968, 1977, 1981)—like that of David Lewis (1966, 1972, 1980)—for the reduction of mental states to brain states. The neurophysiological *realizers* for these causal-functional states would be a matter for empirical science to discover, but by combining these two theories one could see how a completely materialistic account of the mind is possible. Although one need not be a materialist to accept a HO explanation of consciousness, HO theories (as I see them) provide a nice accompaniment to the kind of mind-brain materialism I argued for in chapter 2. Given that we have some intuitive understanding of mental states (like thoughts) independent of the problem of consciousness, HO theories promise a mentalistic reduction of consciousness. Once we have reduced consciousness to more tractable mental states, one could then, in turn, adopt whichever mind-brain account of mental states they see fit to round out the picture.

Let me now turn to the HOT theory itself. The leading principle behind the HOT theory, like all other HO theories, is that a mental state is conscious only if one is, in some suitable way, conscious of that state. This is what we can call the *transitivity principle*. According to this principle, to

be *conscious of* something, or *transitively conscious of* something is to be "in a mental state whose content pertains to that thing" (Rosenthal 1997, 737). For Rosenthal, it is to have a *higher-order thought* of that state. Conscious states themselves, according to this theory, are always *intransitively conscious*. This is what Rosenthal calls *state consciousness* and it is what we seek to explain. State consciousness is a property only of mental states. It is the property of being conscious that some mental states have and others lack. Rosenthal distinguishes state consciousness from both *transitive* consciousness and *creature* consciousness (see 1993c). What we want, Rosenthal argues, is "an account of what it is for mental states to be intransitively conscious on which that property is relational and not all mental states are conscious" (1997, 737). Rosenthal's way of doing this is to explain intransitive or state consciousness in terms of transitive consciousness. And he does this by giving a HOT, as opposed to a HOP, account of transitive consciousness.[20]

A higher-order thought, or HOT, is a thought about some mental state. The core idea of the HOT model is that a mental state is a conscious state when, and only when, it is accompanied by a suitable HOT. And a *thought*, according to Rosenthal, is "any episodic intentional state with an assertoric mental attitude" (1993b, 913 fn.2). Roughly, then, the HOT hypothesis states that a mental state is conscious "just in case one has a roughly contemporaneous thought to the effect that one is in that very mental state" (1993d, 199). This statement of the hypothesis, however, still needs some further restrictions. According to the HOT theory, we must also specify that our transitive consciousness of our mental state "relies on neither inference nor observation . . . of which we are transitively conscious" (1997, 738). One's HOT, that is, must be *noninferential*. This restriction is needed to exclude cases in which one has a thought that one is in a mental state because of the testimony of others, or because one has observed one's own behavior.

We are now in a better position to state the core idea behind the HOT theory of consciousness. We can say that the HOT must be an *assertoric, noninferential, occurrent propositional state*. And we can explain state consciousness as follows: a mental state is a conscious state when, and only when, we have an *assertoric, noninferential, occurrent thought* to the effect that one is in that very mental state.

The HOT theory says that what makes a mental state conscious is the presence of a suitable HOT directed at it. The fact that when we are in conscious states we are typically unaware of having any such HOTs is no objection, for the theory actually predicts that we would not be. Since a mental state is conscious only if it is accompanied by an assertoric, noninferential, occurrent HOT, that HOT will not itself be a conscious thought unless one has a third-order thought about the second-order thought; and Rosenthal points out that this rarely happens. We are conscious of our HOTs only when those thoughts themselves are conscious, and it is

rare that they are. In the rare cases in which this does happen, we would be *introspectively* conscious of being conscious of one's mental states. Hence:

> Most of the time, our mental states are conscious in an unreflective, relatively inattentive way. But sometimes we deliberately focus on a mental state, making it the object of introspective scrutiny. The HOT model readily explains such introspective consciousness as occurring when we have a HOT about a mental state and that HOT is itself a conscious thought. (Rosenthal 2002c, 242)

Introspective consciousness, according to the HOT theory, is the special case of conscious states in which the accompanying HOT is itself a conscious thought because it is the object of a yet higher-order (or third-order) thought.

It is important to note, however, that the HOT theory is *not just* an account of introspective consciousness. Some theorists, I believe, have made this mistake (e.g., Papineau 2002, ch.7).[21] At bottom, the HOT theory explains state consciousness in terms of transitive consciousness or HOTs. But our HOTs, in normal non-introspective conscious experience, are usually nonconscious. As Rosenthal (1986, 1997) and others have pointed out, one can be aware of an experience (via a nonconscious HOT) without introspectively thinking about that experience (see also Gennaro 1996, 16-21; 2003). When one introspects, on the other hand, one's conscious focus is directed back into one's mind, and this will involve a third-order thought (which itself is nonconscious) directed at the HOT. When one is *introspectively aware*, one is conscious of being conscious. The HOT theory has the resources to easily account for this, but one should not confuse the theory with a theory simply of introspective consciousness.

On Rosenthal's account, then, state consciousness turns out to be non-intrinsic and relational. Since a mental state is a conscious state when, and only when, it is accompanied by a suitable HOT, no mental state is essentially conscious. In fact, one of the merits of the HOT model is that it requires that no mental state is essentially conscious. Rosenthal's theory requires that consciousness is a contingent property of mental states, for any mental state that is the object of a HOT presumably need not have been. A mental state is a conscious state only when it is accompanied by a HOT, and is unconscious otherwise. This account fits very nicely with what I have been arguing here. In particular, the HOT model recognizes the existence of unconscious mental states, something Cartesian theories fail to do. This allows us a way of understanding all the interesting research on the automaticity of higher mental processes and the adaptive unconscious discussed in the previous chapter. Reference to nonconscious mental states can be made, according to the theory, on the basis of

the causal role those states play with regard to behavior, shifts in attention, interaction with other mental states, and the like.

Before moving on, there are a few additional aspects of the theory that are worth pointing out. For one, on Rosenthal's account our HOTs should be understood as *extrinsic* to (i.e., entirely distinct from) its target mental state. This differs from Gennaro (1996), for example, who argues that, when one has a first-order conscious state, the HOT is better viewed as *intrinsic* to the target state, so that we have a complex conscious state with parts. He calls this the "wide intrinsicality view" (WIV). I believe Rosenthal's extrinsic account is preferable. For one, it allows a more plausible explanation of cases of misrepresentation and confabulation. If one were to view the HOT as intrinsic to the target states, then it would be hard to explain cases where there is a divergence between the mental states we are in and our awareness of those states. As I argued before, consciousness does not provide us with infallible knowledge of our mental states.[22]

Rosenthal's theory should also be viewed as an *actualist* HOT theory. That is, what makes a mental state conscious, according to Rosenthal, is that it is the object of an *actual* HOT directed at the mental state. This differs, for example, from Carruthers's (1996, 2000) *dispositional* HOT theory, which holds that the HOT need not be actual, but can instead be a dispositional state of some kind. Against the dispositional view, Rosenthal writes, "Simply being disposed to have a thought about something does not result in my being conscious of that thing. For a state to be conscious, my HOT must actually occur" (2002c, 243; also see 2004b). I agree here with Rosenthal, and hence will be appealing to the *extrinsic-actualist* version of the HOT theory in what's to follow.

Furthermore, according to Rosenthal's version of the theory, one should not view the first-order (or target) state as the *only* or *primary* cause of its accompanying HOT (see 1993a). Our HOTs, according to Rosenthal, are sometimes caused by their target states, but there can be other factors that figure in causing it as well. Given that mental states often occur without any accompanying HOT, it cannot be that the states are, by themselves, causally sufficient to produce HOTs; instead "other mental occurrences must enter into the aetiology" (2002c, 245). These other mental occurrences could include our expectations, focus and attention, unconscious monitoring of the environment, and connection to other mental states (both conscious and unconscious). According to the HOT theory, then, which states become conscious is a complex matter determined by a number of different factors. Sometimes the target state will be implicated in causing the accompanying HOT, other times the HOT will be caused by its relation to states other than the target state (both conscious and unconscious).

5.4 MISREPRESENTATION AND CONFABULATION

Since HOTs are extrinsic to lower-order mental states, we can now better understand how misrepresentation and confabulation can occur. HOTs can sometimes represent mental states fully and accurately, but they can also *under-represent*, *misrepresent*, or even *confabulate* those states. As Rosenthal writes: "Typically we see things accurately, and it's also likely that consciousness ordinarily represents correctly what mental states we are in. But misrepresentation of such states can happen (see, e.g., Nisbett & Wilson 1977), and it is an advantage of a higher-order theory that it accommodates such occurrences" (2004b, 35). The HOT theory, I believe, is particularly well suited to explain cases in which the way a conscious state appears differs from the way it actually is. According to the theory, whatever the actual character of a mental state, that state, if conscious, is conscious in respect of whatever mental properties one's HOT represents the state as having. So if our HOT represented us as being in a sensory state with such-and-such a sensory quality, for example, then we would be conscious of that state as having that quality. But since we are dealing with a representational relation between two states, the possibility of misrepresentation always exists. Our HOTs, that is, can fail to represent their targets accurately or fully.

Perhaps the most common form of misrepresentation happens when our HOTs under-represent their target states. As Rosenthal likes to point out, a particular mental state need not be conscious in respect of all of its mental properties. One may be aware, for example, of a throbbing pain only as painful, and not also in respect of its throbbing qualities. Or one may be aware of a sensation of red not in respect of its particular shade, say magenta, but simply as red. According to Rosenthal:

> When one consciously sees something red, one has a conscious sensation of red. The sensation, moreover, will be of a particular shade of red, depending on what shade of red one sees. But, unless one focuses on that shade, one typically isn't conscious of the red in respect of its specific shade; the sensation is conscious only as red of some indeterminate shade. (2002c, 245)

In cases like this, we can say that our HOTs misrepresent the target states they are about since what we are subjectively conscious of does not capture the full nature of those states. Sometimes our HOTs will represent their targets in coarse-grained ways (i.e., *red of some indeterminate shade*), other times—as when one focuses on the experience—the content of our HOTs will be more fine-grained (i.e., *magenta*).

This kind of misrepresentation is probably common with mental states that monitor the environment. Take, for example, the cocktail party effect. When we are at a cocktail party engrossed in a conversation, the other conversations going on around us typically turn into background

noise. In a case like this, we can say that our HOTs represent those other conversations as indistinguishable chatter. This does not mean, however, that our auditory states possess only those properties we are aware of. In fact, our auditory states continue to monitor the environment looking for relevant information. If, for example, our name were to come up in one of those other conversations our attention would quickly shift to that conversation. This suggests that our auditory states have properties and represent the environment in ways that our HOTs do not make us aware of. Because our focal attention is not on those other conversations but is on the conversation we are in, our HOTs do not fully represent those auditory states. This, then, would be another example where our HOTs misrepresent (in the sense of under-represent) our mental states. When our attention is focused on only one part of our conscious experience, the rest of the environment is being under-represented.[23]

If our HOTs can misrepresent the mental states they are about, might it happen that HOTs sometimes occur in the absence of the relevant target state altogether? I believe that if we allow for the misrepresentation of mental states, we need to also allow for the possibility of confabulation—for it is very difficult, perhaps impossible, to distinguish between a case where one drastically misrepresents the mental state they are having, and where one confabulates the state. How much misrepresentation, for example, should we allow before we say it is a different state? Rosenthal, in fact, argues that the distinction between an absent target and a misrepresented target is in an important way arbitrary:

> Suppose my higher-order awareness is of state with property *P*, but the target isn't *P*, but rather *Q*. We could say that the higher-order awareness misrepresents the target, but we could equally well say that it's an awareness of a state that doesn't exist. The more dramatic the misrepresentation, the greater the temptation to say the target is absent; but it's plainly open in any such case to say either. (2004b, 32)

In addition, Rosenthal argues that from a first-person point of view it would be indistinguishable when we are having a HOT together with a target state and when we are having a HOT without a target. If I am conscious of myself as being in a *P* state, he maintains, "it's phenomenologically as though I'm in such a state whether or not I am" (2004b, 35). This aspect of the HOT theory, then, allows us to explain how cases of confabulation could occur, while at the same time preserving the phenomenological appearance.

With these details in place, we can now explain the kind of confabulation discussed earlier by saying that our HOTs represent us as being in states that we are not actually in (or, if you prefer, that our HOTs drastically misrepresent their target states). When this happens, what we are experiencing from a first-person point of view is not the same as what is otherwise going on mentally. We can consciously report being in a *P* state

when, in fact, we are not in any such state. This allows for the possibility of our lower-order mental states causing our behavior in one way, while our HOTs make us aware of a different causal story. Since it is our HOTs which determine *what it's like for us*, and since our HOTs are extrinsic to their target states, Rosenthal's version of the HOT theory is particularly well suited to explain how this can happen. If, for example, we were to confabulate a want or desire, *P*, to explain a particular action *X* —when in reality the true cause for *X* was *Q*—according to the HOT theory, we would subjectively feel as if we are doing *X* because of *P*. Although our HOTs will usually represent their targets accurately, I think it is a virtue of the theory that it allows for the possibility of misrepresentation and confabulation.

Confabulation (or, if you prefer, drastic misrepresentation) is not limited to intentional states either. I have previously argued that we can even confabulate being in pain (Caruso 2005). Children, for example, dislike going to the dentist because it hurts. Researchers have found, however, that there are two components to children's dental pain, painful sensations and anxiety (see Chapman 1980). And researchers have found that anxiety is often confounded with pain (Schacham and Daut 1981) and accounts for about a third of the variance in assessment of pain (Melzack and Torgerson 1971). In fact, this has led researchers to try to devise treatments for dental pain by treating anxiety. These treatments work by changing the patient's attention and imagery (Shapiro 1982) and by heightening the patient's perceptions of self-control (Chapman 1980; Baron, Logan, and Hoppe 1993; Baron and Logan 1993). Giving patients instructions, for example, to focus on sensory (vs. emotional) stimuli during a root canal procedure was found to significantly reduce self-reported pain among patients who were classified as having strong desire for control and low felt control in dental situations (Baron, Logan, and Hoppe 1993).

These findings support the HOT model's claim that one's expectations and interests can help determine what HOTs one will have. We can hypothesize that these patients experience anxiety or fear, along with the lack of control, and consciously react as though in pain, even when local anesthetic makes it unlikely that such pain could be occurring. In these cases the patient can be said to be misidentifying one state, anxiety or fear, for another, pain. And as it turns out, giving the patients back some sense of control (e.g., providing information and stress inoculation training) actually reduces self-reported pain in many of the patients (Baron and Logan 1993). Similar findings have been found in areas other than dental pain. Peter Staats and colleagues (1998), for example, report findings in which rehearsed positive and negative thoughts—whose content is independent of pain—modify the effects of painful stimuli, both as subjectively reported and by standard physiological measures. He found that self-suggestion and the placebo effect, in which genuine assertoric

HOTs presumably occur, significantly altered participants' pain threshold, pain tolerance, and pain endurance.

We can say then that according to the HOT theory: (a) it is the HOT that determines *what it's like* for us; (b) the content of our HOTs can be influenced by factors other than the target state (e.g., the need for control, anxiety, expectations, social preconceptions, idiosyncratic beliefs, conceptual resources, etc.); and (c) our HOTs can misrepresent or confabulate the mental states we are in.

Thus far I have only discussed the possibility of error occurring between our HOTs and their targets, but presumably error can also occur at the level of introspection. Since introspective consciousness, according to the HOT theory, is also the product of a representational relation between two mental states—i.e., when we are introspectively conscious, our HOT is itself the target of a yet higher-order (or third-order) thought—error can happen here too. Some may even find error at this level easier to accept than at the earlier level (e.g., Schooler and Schreiber 2004). At least three independent sets of investigators have recently offered compelling arguments for how discrepancies can occur between our conscious states and our introspective awareness of those states (Schooler 2000, 2001, 2002a, 2002b; Lambie and Marcel 2002; Jack and Shallice 2001; Jack and Roepstorff 2002; for a review, see Schooler and Schreiber 2004).

The approach taken by Jonathan Schooler, for example, focuses on dissociations between consciousness and what he calls *meta-consciousness* (what I am calling introspective consciousness). Schooler and Schreiber describe such dissociations as follows:

> The basic idea underlying the distinction between consciousness and meta-consciousness is simply that individuals often have experiences without necessarily explicitly introspecting about them. As a consequence introspection can fail for two very general reasons. First, introspection may not be invoked. Such introspective failures are what Schooler (2002b) refers to as 'temporal dissociations' in which experience occurs in the absence of meta-awareness. . . . Second, introspection can fail because, in their attempt to characterize an experience, individuals may distort it. [These introspective difficulties] can be characterized as 'translation dissociations' between consciousness and meta-consciousness. Some of the sources of these translation dissociations may result from processes associated with *detection*, *transformation*, and *substitution*. (2004, 31)

So just as our HOTs can distort or misrepresent their target states, our third-order thoughts can likewise distort our conscious mental states. According to Schooler, substitutions, transformations, and difficulties in detection can occur when we attempt to introspect. Anthony Marcel has likewise found that, "Attending to one's experience, introspecting, changes the content, nature and form of the experience" (2003, 179). And

Lambie and Marcel (2002) have argued that introspection can *influence its object*, *create its object*, and *distort its object*.

In terms of the HOT theory, when we introspect and attend to our own conscious experience—whether to examine it, remember it, or report it—our third-order representational states can be influenced by expectations and motivations (just like our second-order HOTs can) thereby distorting the original experience. As Schooler and Schreiber argue, "Introspections may sometimes go awry because the information accessed is not the record of actual experience" (2004, 33). It's possible that instead of retrieving the memory of conscious experience, "people may instead bring beliefs to meta-awareness without realizing that the contents of meta-consciousness may diverge substantially from what was experienced in consciousness" (2004, 33). Expectations, motivations, and idiosyncratic beliefs are among the forces involved in this type of error. Introspection, especially when dealing with introspective reports on behavior, is often more a matter of retro-diction or inference than direct introspection. In fact, many cases of confabulation are probably the result of attempts to introspect mental causes that are no longer occurrent or were unconscious in the first place. When we turn inward in an attempt to introspect why we just behaved as we did, or why we just made the evaluation or judgment we just made, we do not introspect the original mental causes, we introspect a conscious memory. Such introspection is susceptible to influence by internal and external pressures. The content of our third-order state can be influenced, that is, by social expectations and personal idiosyncratic theories.

Some forms of confabulation may therefore be better explained as errors occurring at the level of introspection. I need not make any principled claims, however, about which kinds of errors occur at the level of consciousness and which occur at the level of introspection. Discerning at which level an error occurs might be extremely difficult to do. My interest here is simply to establish that consciousness is neither transparent nor infallible.

5.5 WHAT THE HOT THEORY TELLS US ABOUT FREE WILL

What does the HOT theory tell us about free will? For one, it tells us that Cartesian and libertarian accounts of consciousness, accounts which fit more comfortably with our folk-psychological beliefs about free will, are ill conceived. What the growing research on automaticity and the adaptive unconscious shows us, and what the HOT theory is able to explain, is that: (1) not all mental states are conscious states; and (2) we do not have infallible knowledge of all our higher-level mental states and processes. The HOT theory provides an account of consciousness that complements the research on automaticity and the adaptive unconscious. It explains

why certain mental states are conscious and not others. It also explains, by revealing the relational nature of consciousness, how we can misrepresent or confabulate the mental states we are in.

As we witnessed in chapter 4, higher-level processes (such as goal directed behavior) can occur completely without conscious involvement, and thus automatically. Bargh's (1990) *auto-motive model* of environmentally driven, goal directed behavior, explains how this can happen. The auto-motive model assumes that external events can trigger goals directly, without an explicit conscious choice, and that they then operate without the person knowing it. This fits with the findings of Aarts and Dijksterhuis (2000), who conjecture that habitual behaviors are automatically linked not to relevant environmental events per se but rather to the mental representations of the goal pursuits they serve. Hence, when a goal is unconsciously activated, the habitual plan for carrying out that goal can automatically be activated as well; without need of conscious planning or selection. This can happen when environmental features become automatically associated with the top level or trigger of the goal structure—the same internal representation that is presumably activated by conscious will (see Bargh 1990; Chartrand and Bargh 1996; Bargh and Ferguson 2000).

On the HOT model, such cases of automatic goal-directed behavior would be explained in terms of unconscious first-order mental states. The environment unconsciously and automatically causes in us a first-order mental state—in this case, a mentally represented goal—which carries out its action plan without conscious processing or selection. These first-order mental states remain unconscious since they are not accompanied by HOTs. As Bargh and Ferguson argue:

> Theoretically, this is possible if one assumes that goal representations behave by the same rules as do other mental representations and develop automatic associations to other representations that are frequently and consistently active at the same time (i.e., Hebb's, 1949, principle of contiguous activation; see Shiffrin & Dumais, 1981; Shiffrin & Schneider, 1977). Thus, if a person consistently chooses to pursue the same goal within a given situation, over time that goal structure becomes strongly paired with the internal representation of that situation (i.e., the situational features). Eventually, the goal structure itself becomes active on the perception of the features of that situation. (2000, 934)

Thus, the environment itself could directly activate a first-order unconscious goal as part of the preconscious analysis of the situation. The goal would then operate in the same manner—without the individual knowing it—as when put into play consciously.

Bargh asserts that this can happen not only with habitual behaviors, but also with novel and nonhabitual behaviors (see Bargh 1990; Chartrand and Bargh 1996; Bargh and Ferguson 2000). And the work of Goll-

witzer (1993, 1999; Gollwitzer and Brandstätter 1997) seems to support this claim. All of this research is threatening to free will as normally conceived, for actions that are performed automatically and unconsciously cannot occur freely.

Since the HOT theory is compatible with the automaticity of higher mental processes—especially Bargh's auto-motive model—it too is threatening to free will. Nonconscious mental states have causal roles that are not controlled by our conscious will and these causal roles can be triggered by the environment via a mentally represented goal. Such automatic behavior has traditionally been the paradigm of unfree behavior—for it is commonly defined as unwilled, unintentional, and unaware. According to this traditional perspective, then, "the complexity and the abstract, protracted nature of the kinds of mental processes and social behavior that social-cognition research has recently discovered to operate and occur without conscious, aware guidance have bestowed an unprecedented legitimacy to the traditional conception of determinism" (Bargh and Ferguson 2000, 926). Hence, unlike Cartesian accounts which wrongly deny the existence of unconscious mental states, the HOT theory is compatible with the determination of human behavior by unconscious, automatic processes.

Bargh and Ferguson (2000) point out, however, that the traditional conception of determinism in cognitive and social-cognitive science is inappropriately constrained by the equation of determination with the lack of conscious awareness, choice, and guidance of the process. This constraint gives the impression that consciously mediated acts might be freely willed (i.e., nondetermined). But as Bargh and Ferguson point out:

> [A]lthough the growing social-cognitive evidence of the degree to which higher mental processes can proceed nonconsciously is consistent with the traditional determinist position, by showing that these processes do not require an intervening act of conscious will to occur, it should not be concluded from this that those processes that require conscious or controlled processes (such as those involving temporary and flexible use of working memory; see E.E. Smith & Jonides, 1998) are any less determined. (2000, 926)

Bargh and Ferguson maintain, as I do, that those processes and behaviors that do entail an act of conscious choice are *equally* determined. They write:

> As scientists studying human behavior and the higher mental processes, we reject the thesis of free will as an account of the processes that require conscious control (see also Prinz, 1997). Instead, we embrace the thesis that behavior and other responses are caused, including a person's choices regarding those responses; every deliberation, thought, feeling, motivation, and impulse, conscious or nonconscious, is (often multiply) caused. (2000, 926)

Hence, the presence of conscious control may be a *necessary* condition for free will—as witnessed by our earlier examination of the libertarian position—but it is by no means a *sufficient* condition. Both conscious and nonconscious processes are causally determined according to my account. There is absolutely no reason, then, "to invoke the idea of free will or a nondetermined version of consciousness as a causal explanatory mechanism in accounting for higher mental processes in humans" (Bargh and Ferguson 2000, 939).

There are essentially three reasons for rejecting a nondetermined account of consciousness. For one, as we've already seen, there are insurmountable problems facing libertarian accounts of freedom. Those problems extend to libertarian accounts of consciousness (i.e., nondetermined versions of consciousness). The brain functions according to the principles of classical physics, not quantum physics. Secondly, psychology, cognitive science, and the social-cognitive approach to higher mental processes adopt a deterministic stance toward psychological phenomena (see Bargh and Ferguson 2000; Amsel 1989; Bargh 1997; Barsalou 1992; Zuriff 1985). If one wishes to understand these processes, one should give appropriate due to the deterministic presumptions that underlie such investigation.[24] Thirdly, according to the HOT theory, conscious processes are no different in kind from unconscious processes. Consciousness, according to the HOT theory, is a matter of two mental states being in a certain relation. Although this requires more cognitive energy and resources, one should not think that consciousness operates according to some ontologically distinct set of laws. I therefore second Bargh and Ferguson when they say, "It seems undeniable that conscious processes are themselves causal agents within the same deterministic framework as nonconscious processes" (2000, 939). Conscious and nonconscious processes presumably act in concert with one another and with stimuli outside of our bodies according to the laws of classical physics.

The HOT theory, then, allows us to see that the introspective phenomenology of freedom (at least to the extent that we have examined it here) is misleading. One of the reasons why we impart so much power to the *conscious will*, I maintain, is that we mistakenly believe that we are transparently aware, in an infallible way, of the workings of the mind—particularly the higher-level workings of the mind. It would stand to reason that if one believed that they were infallibly aware of all their mental states, *and one also remained unaware of all the unconscious mental processes that go into guiding judgment, choice, and behavior,* that they would believe the conscious will had far more power than it actually does. But now that we see that a great deal of mental activity is controlled by unconscious mental states, and that we can also misrepresent and confabulate the states we are in, the libertarian argument from introspections loses its force. One may begin to wonder at this point: What function, or functions, does consciousness actually have? If unconscious processes can

control sophisticated behavior, perhaps the causal efficacy of consciousness is itself an illusion (just like our subjective feeling of freedom is an illusion)? To that question I now turn.

5.6 ON THE FUNCTION OF CONSCIOUSNESS

There are many divergent theories on the function, or functions, of consciousness (cf. Baars 1988; Blakemore and Greenfield 1987; Marcel and Bisiach 1988; Velmans 1991; Dehaene and Naccache 2001: Rosenthal 2008). Some of these theories impart a major causal role to consciousness while others argue that consciousness plays little or no role in information processing and higher-level functioning. Over the last two chapters a picture has emerged of a set of pervasive, adaptive, and sophisticated mental processes that occur largely outside conscious awareness. Some theorists have even taken this growing research to the extreme, claiming that the unconscious mind does virtually all the work and that the causal efficacy of consciousness may be an illusion (Huxley 1898; Velmans 1991; Wegner and Wheatley, 1999; Wegner 2002). Although I think there is a danger of going too far here, the role that consciousness plays in causing behavior is probably much smaller than previously believed.

Libet, for example, has shown that the initiation of a spontaneous voluntary act appears to begin in the brain unconsciously, well before one is even aware of the intention to act (Libet et al. 1983).[25] He found that spontaneous voluntary acts are preceded by a specific electrical charge in the brain, a "readiness potential" (RP), that begins 550 msec. before the act. Human subjects, however, only became aware of the intention to act 350-400 msec. after RP starts. It would seem, then, that the voluntary process is therefore initiated unconsciously. Libet argues, however, that even though the conscious awareness of an intention to act comes only after RP, it can still play a role in the final outcome. Since it comes 350-400 msec. after RP, but 200 msec. before the muscle is activated, it can still exercise a "veto" function. Hence, for Libet, "Potentially available to the conscious function is the possibility of stopping or vetoing the final progress of the volitional process, so that no actual muscle action ensues" (1999, 556). On this proposal, conscious will could thus affect the outcome of the volitional process even though the latter was initiated by unconscious cerebral processes. If Libet is correct, and there is much dispute about this, one function of consciousness, at least in spontaneous voluntary action, is its veto power.[26] This, of course, imparts to consciousness a causal function, thereby avoiding epiphenomenalism, but the picture that emerges is a relatively limited one.

This limited role also seems to fit with what we know about the adaptive unconscious. The adaptive unconscious plays a major executive role in our mental lives. It gathers information, interprets and evaluates it,

and sets goals in motion, quickly and efficiently (Wilson 2002). For example, in the Bechara et al. (1997) card game study, we saw that people can figure out which decks had the best payoffs quickly and nonconsciously, without being able to verbalize why they preferred decks C and D. Such ability affords us an invaluable advantage in everyday life. It appears "Our conscious mind is often too slow to figure out what the best course of action is, so our nonconscious mind does the job for us" (Wilson 2002, 36). What was once thought of as the "proper work" of consciousness (e.g., reasoning, evaluation, judgment) can be, and often is, performed nonconsciously. But once we acknowledge that people can think unconsciously in quite sophisticated ways, questions arise about the relation between conscious and nonconscious processing. If we can reason, evaluate, and make judgments without consciousness, what function, if any, does consciousness have?

Theorists have suggested a number of possibilities. Consciousness has been thought to be necessary for the analysis of novel stimuli or novel stimulus arrangements (e.g., Posner and Snyder 1975; Bjork 1975). Some theorists, like Mandler (1975, 1985), have argued that consciousness allows us to choose amongst competing input stimuli. Other theorists have assumed that conscious processing is necessary for a stimulus to be remembered (James 1890; Underwood 1979; Waugh and Norman 1965) and for the production of anything other than an automatic, well-learnt response—e.g., for a voluntary response that is flexible or novel, or for a response that requires monitoring or planning (Romanes 1895; Mandler 1975, 1985; Shiffrin and Schneider 1977; Underwood 1982). Although consciousness may play a role in these functions, one should be careful not to overstate the case. One should not, for example, claim that consciousness is necessary or essential for these functions. Recent research on automaticity has shown that many of these functions can be performed nonconsciously.

Proving that consciousness is necessary for a particular function is extremely hard to do (see Rosenthal 2008; Flanagan and Polger 1995, 1998). Owen Flanagan, for example, has introduced something he calls the thesis of conscious inessentialism. This is the thesis that for any intelligent activity, i, performed in any cognitive domain d, even if we do i with conscious accompaniments, i can in principle be done without these conscious accompaniments (1992, 5). Although some have objected to this thesis (Dennett 1995), it is important to point out that conscious inessentialism is a very weak claim. As Polger and Flanagan describe the thesis:

> It is a claim about the mere possibility of some creature that can behave as we conscious beings do, but without consciousness. One way this might be true, of course, is if consciousness is an epiphenomena. But that is not the only way. It may be the case that consciousness is causal-

> ly efficacious, but that the functions that it performs can be accomplished—at least in principle—by non-conscious mechanisms. So conscious inessentialism is compatible with a thorough-going naturalism about the mechanisms and subvenient basis of consciousness, and with a variety of claims about the causal efficacy of consciousness for us. According to this view, consciousness is a mechanism by which some important cognitive functions are performed in human beings. But the fact that we perform these functions consciously is contingent. (1999, 2)

Conscious inessentialism, then, is consistent with consciousness being causally efficacious. It only creates a challenge for those who wish to maintain that a particular function of consciousness is necessary or essential. For that reason, I think it is best to avoid claims of necessity altogether. It may well be that the conscious cognitive functions in human beings, whatever those may be, are contingent.[27] For my purposes, I need not take a stand on this.

All I wish to argue here is that consciousness is not epiphenomenal. I have all along maintained three main theses: (1) that our mental states (both conscious and unconscious) do, at least sometime, play a causal role in our choices and actions; (2) that these causal roles are completely determined by conditions we, ourselves, have no ultimate control over; and (3) although consciousness is not epiphenomenal, our conscious feeling of freedom is deeply deceptive.

Hence, the thesis of conscious inessentialism does not create a problem for what I wish to argue. It does, on the other hand, create a problem for libertarians like O'Connor who wish to argue that a necessary function of consciousness is its agent-causal capacity. As we witnessed in the first section of this chapter, libertarians *must* argue that one essential function of consciousness is that it somehow exercises an agent-causal power. If—even if only theoretically—all mental functions can be performed nonconsciously, the need for an agent-causal function to explain human behavior would vanish. If high-level voluntary action can be performed without conscious control and guidance, it is unnecessary to posit an agent-causal function for such behaviors. Hence, one of the biggest challenges for the libertarian—a challenge I do not believe they can meet given everything we know about human cognition—is to prove that agent-causation is an essential function of consciousness.

I believe the most appropriate view of the role of consciousness is somewhere between the extremes. Although consciousness lacks the kind of libertarian control we traditionally assume it has, it is also far from epiphenomenal. Following Timothy Wilson, I reject both the analogy of consciousness-as-chief-executive of the mind and the view that consciousness is epiphenomenal (2002, 46-47). I agree with Wilson that "we know less than we think we do about our own minds, and exert less control over our own minds than we think. And yet we retain some ability to influence how our minds work" (2002, 48). My position is that

consciousness performs a number of important functions—functions that unconscious and automatic processes are not as good at (e.g., dealing with variable and novel stimuli, making long-range action plans)—but that these functions are more limited than we typically think. I maintain that conscious mental states are causally efficacious, just as nonconscious mental states are, but that the cognitive energy and resources required to generate HOTs means that conscious processes are often slower than nonconscious processes and often follow such processes.

As Wilson points out, the adaptive unconscious is extremely good at detecting patterns in the environment and evaluating them. Such a system has obvious advantages, but it also comes with a cost: "the quicker the analysis, the more error-prone it is likely to be" (2002, 50). Wilson speculates, "It would be advantageous to have another, slower system that can provide a more detailed analysis of the environment, catching errors made by the initial, quicker analysis" (2002, 50). This, he argues, is the job of conscious processing. If this is correct, consciousness acts more like an after-the-fact checker and balancer than as an executive controller of everything mental. This fits in with Libet's suggestion that consciousness has a veto function over spontaneous voluntary action but is not itself the initiator of the action. It also fits with Joseph LeDoux's (1996) suggestion that humans have a nonconscious "danger detector" that sizes up incoming information before it reaches conscious awareness. As Wilson describes this nonconscious danger detector:

> If it determines that the information is threatening, it triggers a response. Because this nonconscious analysis is very fast it is fairly crude and will sometimes make mistakes. Thus it is good to have a secondary, detailed processing system that can correct these mistakes. Suppose that you are on a hike and suddenly see a long, skinny, brown object in the middle of the path. Your first thought is "snake!" and you stop quickly with a sharp intake of breath. Upon closer analysis, however, you realize that the object is a branch from a small tree, and you go on your way. (2002, 5)

According to Wilson and LeDoux, we first perform an initial, crude analysis of the stick nonconsciously, followed by a more detailed conscious analysis. In terms of the HOT theory we can say that unconscious mental states are continually processing and analyzing information about the environment, but when something is deemed important additional cognitive resources are brought to bear causing a HOT of that state.

I do not mean to suggest, however, that consciousness is simply a back-up system for the adaptive unconscious, or that its only function is to veto what unconscious processes have already set in motion. Consciousness, I maintain, also helps facilitate long-term memory formation, plays an important role in non-spontaneous decision-making, provides focal-attention to help prioritize and recruit subgoals and functions, and

serves a metacognitive or self-monitoring function (see Baars 1988). Take the encoding of long-term memory for example. On intuitive grounds, as Max Velmans writes, "it is difficult to envisage how, without consciousness, one could update long-term memory, for if one has never experienced an event, how could one remember it? How could an event which is not part of one's psychological present become part of one's psychological past?" (1991, sec. 4.2). I think it would be relatively uncontroversial to claim that one important function of consciousness is the role it plays in encoding long-term memory.[28] Although preconscious (or nonconscious) contents may influence the way the contents of consciousness are interpreted and consequently remembered (see Lackner and Garrett 1972; MacKay 1973), preconscious contents themselves do not usually enter into long-term memory.[29] Hence, as Velmans points out, "It is the accepted wisdom (backed by numerous experiments) that unless preconscious contents are selected for focal-attentive processing and enter consciousness, they are quickly lost from the system (within 30 seconds)" (1991, sec. 4.2). One should be very careful, however, not to confuse the contents of consciousness and long-term memory with the processes which encode information, transfer it to long-term memory, and search and retrieve it. Such processes are not under our conscious control and are inaccessible to introspection.

In addition to the role consciousness plays in encoding long-term memory, it's also likely that it plays an important role in representing and adapting to novel and significant events. Whereas the adaptive unconscious is fast, automatic, and effortless, it is also rigid. Consciousness processes, on the other hand, though slower and effortful, are much more flexible. According to Bernard Baars, "the most fundamental function [of consciousness] is . . . the ability to optimize the trade-off between organization and flexibility." He argues that, "Organized responses are highly efficient in well-known situations, but in cases of novelty, flexibility is at a premium" (1988, 348). It's useful, then, that the adaptive unconscious make "canned" solutions available automatically in predictable situations, but that the cognitive system be capable of combining all possible knowledge sources in unpredictable circumstances. Although the adaptive unconscious allows us to perform many behaviors quickly, effortlessly, and automatically, consciousness allows for plasticity and flexibility when it comes to novel situations. As Baars points out, "as soon as we flag some novel mental event consciously, we may be able to recruit it for voluntary action" (1988, 352).

According to the HOT theory, consciousness adds a spotlight to mental processes making certain mental states "light up." In so doing, it brings focal-attention to those states and processes allowing the mind to use that information in novel ways. It is important to note, however, that the kind of conscious control one has in such situations is nothing like the kind of conscious control assumed by libertarians. Baars argues, for ex-

ample, that conscious goals can help recruit novel subgoals and motor systems to organize and carry out mental and physical actions, but that such recruitment is not itself a matter of conscious control—since "conscious goal-images themselves are under the control of unconscious goal contexts, which serve to generate a goal-image in the first place" (1988, 352). Baars further argues, when automatic systems cannot routinely resolve some choice-point "making it conscious helps recruit unconscious knowledge sources to make the proper decision" (1988, 349). Hence, in the case of indecision, we can make a goal conscious to "allow widespread recruitment of conscious and unconscious 'votes' for and against it" (1988, 349). Although consciousness may not control which choice we make, making the indecision or deliberation conscious allows unconscious knowledge and subgoals to work more effectively on the problem. Making a choice-point conscious allows the answer to be searched for unconsciously. In turn, "candidate answers are returned to consciousness, where they can be checked by multiple unconscious knowledge sources" (1988, 352).

If this is correct, although consciousness does not exercise executive control, it can still play a causal role in setting goals and making decisions. Consciousness can serve, that is, as "the domain of competition between different goals, as in indecisiveness and in conscious, deliberate decision" (Baars 1988, 353). When I become conscious of a choice—"Should I finish this chapter now, or should I stop for lunch?"—this allows a coalition of unconscious systems to build up in support of either alternative, as if they were voting one way or another. Once a decision is reached it can be "broadcast" to the rest of the system—i.e., the decision can be made conscious by having a HOT of it—and action can be taken. Although this is a very important function of consciousness, it is far from the kind of conscious control we often assume exists in libertarian and folk-psychological accounts of free will.

In this way consciousness can also act as the theatre of deliberation for long-range decision making and planning about the future. Hence, consciousness can play an active role in choosing a career path, deciding on what classes to take next semester, and whom to marry.[30] As Wilson points out, unconscious processes are more concerned with the here-and-now whereas conscious processes are better suited for the long view (2002, 50-52). Although the adaptive unconscious reacts quickly to our current environment, skillfully detects patterns, alerts us to any dangers, and sets in motion goal-directed behaviors, it cannot anticipate what will happen tomorrow, next week, or next year, and plan accordingly. "Nor can the adaptive unconscious muse about the past and integrate it into a coherent self-narrative" (Wilson 2002, 51). Consciousness, on the other hand, affords us the ability to reflect on the past and to contemplate the future. Having a flexible mental system that can muse, reflect, ponder,

and contemplate alternative futures and connect those scenarios to the past is therefore a great advantage.

I have here only speculated on a few functions consciousness may perform. This is not meant to be an exhaustive list. Baars (1988), for example, lists no less than eighteen different functions. Whatever the final story turns out to be with regard to the functions of consciousness, I believe we will find that consciousness is not epiphenomenal, as some have assumed, but that it is also not the chief executive of the mind either. Current research, I argue, points to something in between. Whereas the adaptive unconscious appears to be a parallel processing system with multiple modules all functioning and working away at the same time, consciousness is more than likely a serial, limited capacity system. It allows us to perform a number of important tasks, no doubt, and it imparts to us advantages that an unconscious system would lack, but it is not the vehicle of libertarian freedom.

NOTES

1. For additional examples of libertarians appealing to introspective evidence as support for their position, see C.A. Campbell (1957, 176-178), O'Connor (2000, 124), and Kane (1996, 147). Thomas Reid goes even further in claiming that the conception of a cause is itself dependent on our introspective consciousness of self-agency. He writes: "It is very probable, that the very conception or idea of active power, and efficient causes, is derived from our voluntary exertions in producing effects; and that, if we were not conscious of such exertions, we should have no conception at all of a cause, or of active power" (Reid 1895, 2:604).

2. Given the libertarian reliance on introspection, Peter Ross has recently argued that "psychological research into the accuracy of introspection is the source of the most powerful empirical constraint for the problem of free will" (2006,134). According to Ross, "The Libertarian claims that the best explanation of our feeling that there are metaphysically open branching paths is that we become aware of an absence of sufficient mental causes. A specific question for research is whether this is the best explanation. If psychologists were to provide an alternative explanation which not only indicates that there are sufficient mental causes even in ordinary cases where our introspection indicates otherwise, but also offers a model explaining the illusion of their absence, this would undermine any . . . libertarianism" (2006, 139).

3. See section 5.6 of this chapter for more on the possible functions of consciousness.

4. Shaun Nichols (2004, 2006a, 2006b) offers this analysis up as one possible psychological explanation of where the belief in libertarian freedom comes from. Nichols ultimately rejects it in favor of another alternative—one based on the notion of moral obligation. I, on the other hand, believe the hypothesis should be given more weight than Nichols. It is my contention that phenomenology plays a much larger role in producing the illusion of libertarian freedom than Nichols acknowledges.

5. To be clear, I am not arguing that this fully explains the illusion of free will. My position is that the *apparent* transparency and infallibility of consciousness contributes to the illusion but is only one component of the overall story. I do not think it is the *whole* story because it does not explain why we also feel the positive power of active freedom and self-determination. Not being aware of deterministic causes may explain why we believe no such causes exist, but it does not fully explain the phenomenology

of agent-causation. To fully explain why we feel free, I believe we will also have to examine the phenomenology surrounding our feeling of "self-causation" and our feeling of "intentional control" over our actions. I will attempt to explain these features of the illusion in the following chapters.

6. Evidence for my proposal can be found in the fact that when the *feeling* of transparency and infallibility is thrown into question for various reasons, we tend to experience a loss or reduction in the feeling of will. When, for example, we become aware of reasons for acting that are different than the ones that appear directly in consciousness, as when we come to infer (e.g., through some kind of causal reasoning) that the best explanation for our action is something other than the one we were consciously aware of at the time (e.g., habit, emotion, addiction, or other unconscious action tendencies), we often experience a diminished sense of freedom. See Wegner (2002) for documented examples of this.

7. The arguments in this section will mirror the reasoning and structure found in Rosenthal (1986, 1997, and 2002c). I do not claim originality here. These arguments are meant to show that the HOT theory is preferable—empirically and explanatorily—over a Cartesian alternative.

8. I have elsewhere called this the "Cartesian assumption" and have presented arguments against it. See Caruso (2005).

9. All Descartes references will be to either, *The Philosophical Writings of Descartes Vols. 1-3*, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch (hereafter referred to as CSM) or to *Oeuvres de Descartes*, eds. Charles Adam and Paul Tannery (hereafter AT). This quote, for example, can be found in CSM, volume II, page 171. References to AT will follow the same pagination.

10. In the *Principles* (I, 9) Descartes also defines "thought" in terms of consciousness or immediate consciousness (AT, VIII:7-8). And in the *First Set of Replies* Descartes says he can "affirm with certainty that there can be nothing within me of which I am not in some way aware" (CSM, II:77; AT, VII:107). He's even more forceful in a letter to Mersenne where he writes: "What I say later, 'nothing can be in me, that is to say, in my mind, of which I am not aware', is something which I proved in my *Meditations*" (CSM, III:165; AT, III:273).

11. For Descartes, sensations, as far as they are mental states, are just a special kind of thought. He believed that sensations were either a special kind of thought and therefore conscious, or bodily states and hence never conscious. So for something to be considered a mental state, for Descartes, it had to be a thought. And since all thoughts are conscious, all mental states must be as well.

12. For more on this issue, see Rosenthal (1986, 2002c).

13. For a full account of how sensory qualities can exist independently of consciousness, see Rosenthal (1991a).

14. Rosenthal (1993c) has introduced a widely acknowledged distinction between *creature* consciousness and *state* consciousness. *Creature* consciousness recognizes that we often speak of whole organisms as conscious or aware. *State* consciousness, on the other hand, recognizes that we also speak of individual mental states as conscious. Explaining state consciousness is the primary focus for most researchers, and it is what I am here seeking to explain.

15. Although Descartes famously entertained the possibility that the content of my mental states may not match reality, he never entertained the possibility that my own mental states may diverge from my conscious awareness of them. Descartes claims in numerous places that one cannot be mistaken about how things *seem* to them to be in consciousness (e.g., CSM, II:19; AT, VII:29). For Descartes, our judgments about our mental states are certain and infallible.

16. See Wilson (2003) for more on the limits of introspective reports.

17. For a review of confabulation, which includes examples from split-brain patients, people suffering from organic amnesia, and people acting out posthypnotic suggestion, see Wilson (2002, ch.5).

18. In their original paper, Nisbett and Wilson (1977) argued that people often make inaccurate reports about the causes of their responses because there is little or no introspective access to higher order cognitive processes. They theorized that when people try to give introspective reports on the causes of their behavior, what they are really doing is making reasonable inferences about what the causes must have been, not giving direct introspective reports of the actual causes. A number of critics accurately pointed out that this thesis was far too extreme (Smith and Miller 1978; Ericsson and Simon 1980; Gavanski and Hoffman 1987), and Wilson has since modified his views (see Wilson and Stone 1985; Wilson 2002). My position is that we often do have direct access to our own mental states (i.e., our higher cognitive processes), but that the way we are aware of such states allows for the possibility of misrepresentation and confabulation. I further maintain that we need to distinguish between ordinary consciousness and introspective consciousness. I believe that error can occur at both levels and that people often conflate the two types of mistakes. My position is more fully spelled out in section 5.4.

19. Because of its hierarchical or iterative nature, Guzeldere (1995) has called such theories "double-tiered" theories.

20. It is important to point out that there is no circularity here, since the transitively conscious state (or HOT) is not normally itself intransitively conscious.

21. For an analysis of this mistake, see Gennaro (2003).

22. See Rosenthal (2004b) for more on why the extrinsic version of the HOT theory is preferable. And see section 5.4 below for an account of how Rosenthal's extrinsic HOT theory explains cases of misrepresentation and confabulation.

23. Rosenthal suggests that our HOTs can also misrepresent our mental states in the opposite direction; they can fill in information. See, for example, Rosenthal (2004b, 35).

24. The field of psychology presupposes determinism. As three leading psychologists have recently put it: "For psychology to make any sense, the universe must be, to some degree at least, predictable. A psychology that doesn't accept causes of behavior or the possibility of prediction is no psychology at all" (Baer, Kaufman, and Baumeister 2008, 4).

25. There are, however, critics of Libet's findings but I will address these in the following chapter. In chapter 6, I will discuss Libet's methodology along with alternative interpretations and criticisms. I will also offer my HOT interpretation of the findings and differentiate it from similar accounts given by Rosenthal (2002a) and Wegner (2002).

26. Although Libet's "veto power" represents a *potential* function of consciousness, it's important to keep in mind that Libet's proposal is much disputed. The best Libet himself can say for it is that it is not absolutely ruled out by the evidence. In fact, there have been a number of criticisms of Libet's claim (see Velmans 2003) and recent empirical findings by Simone Kühn and Marcel Brass (2009) appear to suggest that the decision to "veto" an action is itself determined unconsciously, just as the initiation of spontaneous voluntary actions appears to be.

27. Rosenthal, for example, claims: "Indeed, it is in any case puzzling what evolutionary pressure there could have been for mental states to be conscious, whatever the explanation of their being conscious. Evolutionary pressure on mental functioning operates only by way of interactions that such functioning has with behavior. And mental functioning interacts with behavior solely in virtue of its intentional and qualitative properties. If a mental state's being conscious does consist in its being accompanied by a higher-order state, that higher-order state would contribute to the overall causal role, but this contribution would very likely be minimal in comparison with that of the first-order state. So there could be little adaptive advantage in states' becoming conscious" (2004b, 27; see also 2008). For a different view of the adaptive value of higher-order thoughts, see Rolls (2004).

28. As I said earlier, though, it is important to steer clear of claims of necessity. Although a strong case could be made for the thesis that consciousness is necessary for

the encoding of long-term memory, Velmans (1991) has argued that some studies on hypnosis indicate that information may be able to enter long-term memory and be recalled without first entering consciousness.

29. Velmans (1991) points out, however, that preconscious processing can affect the memory trace of an input stimulus even if that stimulus cannot later be explicitly recognized or recalled. For an example, see Eich (1984).

30. It should be noted, however, that unconscious evaluations, judgments, and goals also factor heavily in such decisions. Although consciousness allows us to look into the future and try to set long-term goals, it cannot go completely against what the rest of the mental system wants. It is constrained, to a large degree, by our unconscious states and processes.