Consciousness, Free Will, and Moral Responsibility

Gregg D. Caruso

[Edited and abbreviated version forthcoming in *Routledge Handbook of Consciousness*. Edited by Rocco J. Gennaro]

You may cite this longer version as:

Caruso, Gregg D., Consciousness, Free Will, and Moral Responsibility (December 21, 2016). Unedited version. Available at SSRN: https://ssrn.com/abstract=2888513

In recent decades, with advances in the behavioral, cognitive, and neurosciences, the idea that patterns of human behavior may ultimately be due to factors beyond our conscious control has increasingly gained traction and renewed interest in the age-old problem of free will. To properly assess what, if anything, these empirical advances can tell us about free will and moral responsibility, we first need to get clear on the following questions: Is consciousness necessary for free will? If so, what role or function must it play? Are agents morally responsible for actions and behaviors that are carried out automatically or without conscious control or guidance? Are they morally responsible for actions, judgments, and attitudes that are the result of implicit biases or situational features of their surroundings of which they are unaware? What about the actions of somnambulists or cases of extreme sleepwalking where consciousness is largely absent? Clarifying the relationship between consciousness and free will is imperative if we want to evaluate the various arguments for and against free will. For example, do compatibilist *reasons-responsive* and *deep self* accounts require consciousness? If so, are they threatened by recent developments in the behavior, cognitive, and neurosciences? What about libertarian accounts of free will? What powers, if any, do they impart to consciousness and are they consistent with our best scientific theories about the world?

In this chapter I will outline and assess several distinct views on the relationship between consciousness and free will, focusing in particular on the following four broad categories:

(1) The first maintains that consciousness is a necessary condition for free will and that the condition can be satisfied. Such views affirm the existence of free will and claim conscious control, guidance, initiation, broadcasting, and/or awareness are essential for free will. Different accounts will demand and impart different functions to consciousness, so this category includes a number of distinct views (see, e.g., Hodgson 2002, 2005, 2008, 2012; Searle 2000, 2001a, b, 2007; see also Baumeister, Mele, and Vohs 2010).

(2) The second category also maintains that consciousness is a necessary condition for free will, but holds that recent developments in the behavioral, cognitive, and neurosciences either shrinks the realm of free and morally responsible action or completely eliminates it. I include here two distinct types of positions: (2a) The first denies the causal efficacy of *conscious will* and receives its contemporary impetus from pioneering work in

neuroscience by Benjamin Libet, Daniel Wegner, and John-Dylan Haynes; the second (2b) views the real challenge to free will as coming, not from neuroscience but from recent work in psychology and social psychology on *automaticity, situationism, implicit bias*, and the *adaptive unconscious*. This second class of views does not demand that *conscious will* or *conscious initiation of action* is required for free will, but rather conscious awareness, broadcasting, or integration of certain relevant features of our actions, such as their morally salient features. It further maintains that developments in psychology and social psychology pose a threat to this consciousness condition (see Caruso 2012, 2015a; 2015b; Levy 2014).

(3) A third class of views simply thinks consciousness is irrelevant to the free will debate. I include here traditional conditional analyses approaches as well as many *deep self* and *reasons-responsive* accounts that either ignore or explicitly reject a role for consciousness. Thomas Pink, for example, writes of the first group:

> [M]any philosophers, especially in the English-language tradition, have taken the view that the question of free will has nothing to do with consciousness. For them the free will problem is about the correct semantic analysis of the expression 'could have done otherwise'; and such an analysis is to be provided simply by considering concepts or sentence meanings, without any reference to consciousness or experience. (2009: 296)

More recently, however, a growing number of contemporary philosophers have explicitly rejected a consciousness condition for free will, focusing instead on features of the agent that are presumably independent of consciousness. Prominent examples include Nomy Arplay (2002), Angela Smith (2005), and George Sher (2009). These philosophers typically rely on everyday examples of agents who appear free and morally responsible in the relevant sense but who act for reasons of which they are apparently unconscious.

(4) The final category of views has played an important role in the historical debate over free will but is a bit orthogonal to the others. Rather than explicitly embracing or rejecting the idea that consciousness is necessary for free will, these views appeal to consciousness, especially the supposed consciousness we have of our own freedom, as evidence for the reality of free will. Such views tend to be libertarian in nature, especially agent-causal, rather than compatibilist, and appeal to the phenomenology of agency as *prima facie* support for the existence of free will (e.g., Campbell 1957: 169; Taylor 1992: 51; O'Connor 1995: 196; Kane 1996).

I. Free Will and Moral Responsibility

Before discussing each of the categories in detail, let me begin with a few definitions. First, the kind of *free will* I will take to be of central philosophical and practical importance in the historical debate is the control in action required for a core sense of moral responsibility. This sense of moral responsibility is set apart by the notion of *basic desert* and is purely backward-looking and non-consequentialist (see Feinberg, 1970; Pereboom, 2001, 2014; G. Strawson, 1994; Fischer, 2007; Caruso and Morris 2016). As Derk Pereboom defines it:

> For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations. (2014: 2)

Understood this way, free will is a kind of power or ability an agent must possess in order to justify certain kinds of desert-based judgments, attitudes, or treatments in response to decisions or actions that the agent performed or failed to perform. These reactions would be justified on purely backward-looking grounds and would not appeal to consequentialist or forward-looking considerations, such as future protection, future reconciliation, or future moral formation.

Second, contemporary theories of free will tend to fall into one of two general categories, namely, those that insist on and those that are skeptical about the reality of human freedom and moral responsibility. The former category includes *libertarian* and *compatibilist* accounts of free will, two general views that defend the reality of free will but disagree on its nature. The latter category includes a family of skeptical views that all take seriously the possibility that human beings do not have free will, and are therefore not morally responsible for their actions in the basic desert sense. The main dividing line between the two pro-free will positions, libertarianism and compatibilism, is best understood in terms of the traditional problem of free will and determinism. *Determinism*, as it is commonly understood, is roughly the thesis that every event or action, including human action, is the inevitable result of preceding events and actions and the laws of nature. The problem of free will and determinism therefore comes in trying to reconcile our intuitive sense of free will with the idea that our choices and actions may be causally determined by impersonal forces over which we have no ultimate control.

Libertarians and compatibilists react to this problem in different ways. Libertarians maintain that if determinism is true, and all of our actions are causally necessitated by antecedent circumstances, we lack free will and moral responsibility. Yet they further maintain that at least some of our choices and actions must be free in the sense that they are not causally determined. Libertarians therefore reject determinism and defend an indeterminist conception of free will in order to save what they believe are necessary conditions for free will—i.e., either the *ability to do otherwise* in exactly the same set of conditions or the idea that we remain, in some important sense, the *ultimate source/originator* of action. Compatibilists, on the other hand, set out to defend a different form of free will, one that can be reconciled with the acceptance of determinism. They hold that what is of utmost importance is not the falsity of determinism, nor that our actions are uncaused, but that our actions are voluntary, free from constraint and compulsion, and caused in the appropriate way. Different compatibilist accounts spell out the exact requirements for compatibilist freedom differently but popular theories tend to focus on such things as reasons-responsiveness, guidance control, hierarchical integration, and approval of one's motivational states.

In contrast to these pro-free will positions are those views that either doubt or outright deny the existence of free will and moral responsibility. Such views are often referred to as skeptical

views, or simply *free will skepticism*. In the past, the standard argument for skepticism was *hard determinism*: the view that determinism is true, and incompatible with free will and moral responsibility—either because it precludes the *ability to do otherwise* (leeway incompatibilism) or because it is inconsistent with one's being the "ultimate source" of action (source incompatibilism). For hard determinists, libertarian free will is an impossibility because human actions are part of a fully deterministic world and compatibilism is operating in bad faith. Most contemporary free will skeptics, however, offer arguments that are agnostic about determinism—e.g., Derk Pereboom (2001, 2014), Galen Strawson (1986/2010), Neil Levy (2011), Richard Double (1991), Bruce Waller (2011), and Gregg Caruso (2012). Most maintain that while determinism is incompatible with free will and moral responsibility, so too is *indeterminism*, especially the variety posited by quantum mechanics (Pereboom 2001, 2014; Caruso 2012). Others argue that regardless of the causal structure of the universe, we lack free will and moral responsibility because free will is incompatible with the pervasiveness of *luck* (Levy 2011). Others (still) argue that free will and ultimate moral responsibility are incoherent concepts, since to be free in the sense required for ultimate moral responsibly we would have to be *causa sui* (or "cause of oneself") and this is impossible (Strawson 1994, 1986).

In addition to these philosophical arguments, there have also been recent developments in the behavioral, cognitive, and neurosciences that have caused many to take free will skepticism seriously. Chief among them have been findings in neuroscience that appear to indicate that unconscious brain activity causally initiates action prior to the conscious awareness of the intention to act (e.g., Libet et al. 1983; Libet 1985, 1999; Soon et al. 2008), Daniel Wegner's (2002) work on the double disassociation of the experience of will, and recent findings in psychology and social psychology on automaticity, situationism, and the adaptive unconscious (e.g., Bargh 1997, 2008; Bargh and Chartrand 1999; Bargh and Ferguson 2000; Wilson 2002; Nisbett and Wilson 1977; Doris 2002).[1] Viewed collectively, these developments suggest that much of what we do takes place at an automatic and unaware level and that our commonsense belief that we consciously initiate and control action may be mistaken. They also indicate that the causes that move us are often less transparent to ourselves than we might assume—diverging in many cases from the conscious reasons we provide to explain and/or justify our actions. These findings reveal just how wide open our internal psychological processes are to the influence of external stimuli and events in our immediate environment, without knowledge or awareness of such influence. No longer is it believed that only "lower level" or "dumb" processes can be carried out non-consciously. We now know that the higher mental processes that have traditionally served as quintessential examples of "free will"—such as evaluation and judgment, reasoning and problem solving, and interpersonal behavior—can and often do occur in the absence of conscious choice or guidance (Bargh and Ferguson 2000: 926). Neil Levy calls this the "automaticity revolution" and it consists in "recognizing the major role that automatic processes play in psychology, and therefore behavior" (2014: 4; cf. Shepherd 2015b).

For some, these findings represent a serious threat to our everyday folk understanding of ourselves as conscious, rational, responsible agents, since they indicate that the conscious mind exercises less control over our behavior than we have traditionally assumed. In fact, even some compatibilists now admit that because of these behavioral, cognitive, and neuroscientific

[1] The literature on *Social Intuitionism* (e.g., Haidt 2001) is also sometimes cited in this regard.

findings "free will is at best an occasional phenomenon" (Baumeister 2008: 17). This is an important concession because it acknowledges that the *threat of shrinking agency*—as Thomas Nadelhoffer (2011) calls it—remains a serious one independent of any traditional concerns over determinism. That is, even if one believes free will and causal determinism can be reconciled, the deflationary view of consciousness which emerges from these empirical findings must still be confronted, including the fact that we often lack transparent awareness of our true motivational states. Such a deflationary view of consciousness is potentially agency undermining and must be dealt with independent of, and in addition to, the traditional compatibilist/incompatibilist debate (see, e.g., Caruso 2012, 2015b; Levy 2014; Nadelhoffer 2011; King and Carruthers 2012; Sie and Wouters 2010; and Davies 2009).

II. Is Consciousness Necessary for Free Will?

Turning now to the relationship between consciousness and free will, the four categories outlined above are largely defined by how they answer the following two questions: (1) Is consciousness necessary for free will? And if so, (2) can the consciousness requirement be satisfied given the threat of shrinking agency and recent developments in the behavioral, cognitive, and neurosciences? Beginning with the first question, we can identify two general sets of views—those that reject and those that accept a *consciousness condition* on free will. The first group includes philosophers like Nomy Arpaly (2002), Angela Smith (2005), and George Sher (2009), who explicitly deny that consciousness is needed for agents to be free and morally responsible. The second group, which includes Neil Levy (2014), myself (Caruso 2012, 2015b), and Joshua Shepherd (2012, 2015a), argue instead that consciousness *is* required and that accounts that downplay, ignore, or explicitly deny a role for consciousness are significantly flawed and missing something important.

Among those who deny that consciousness is necessary for free will are many proponents of the two leading theories of free will and moral responsibility: *deep self* and *reasons-responsive* accounts. Contemporary proponents of deep self accounts, for instance, advocate for an updated version of what Susan Wolf (1990) influentially called the *real self* view, in that they ground an agent's moral responsibility for her actions "in the fact…that they express who she is as an agent" (Smith 2008: 368). According to deep self (or real self) accounts, an agent's free and responsible actions should bear some kind of relation to the features of the psychological structure constitutive of the agent's *real* or *deep* self (Arpaly and Schroeder 1999; Arpaly 2002; Wolf 1990). As Faraci and Shoemaker describe such views:

> [T]he basic idea has been to identify a subset of an agent's motivating psychological elements as privileged for self-determination and responsibility, such that as long as one's actions are ultimately governed by this subset, they count as one's own and thus render one eligible for responsibility-responses to them. (2010: 320; as quoted by Shepherd 2015a).

Deep self theorists typically disagree on which psychological elements are most relevant, but importantly none of them emphasize consciousness. In fact, some explicitly deny that expression of who we are as agents requires that we be conscious either of the attitudes we express in our

actions or the moral significance of our actions (see, e.g., Arpaly 2002; Smith 2005). Deep self accounts therefore generally fall into the third category identified in the introduction.

Reasons-responsive accounts also tend to dismiss the importance of consciousness. According to John Martin Fischer and Mark Ravizza's (1998) influence account, responsibility requires not *regulative* control—actual access to alternative possibilities—but only *guidance* control. And, roughly speaking, an agent exercises guidance control over her actions if she would recognize reasons, including moral reasons, as reasons to do otherwise, and she would actually do otherwise in response to some such reasons in a counterfactual scenario. But, as describes, such accounts impart no significant role to consciousness:

> According to Fischer and Ravizza's well-known version of this view, an agent is morally responsible for an action only if that actions is produced via a mechanism that both recognizes and reacts—in a sufficiently flexible way, and typically via action—to reasons for action (1998). There is much to like about this kind of view. But on the face of it, a Reasons-Responsive View highlights considerations orthogonal to consciousness. It is true that a connection between consciousness and reasons-responsiveness is sometimes *assumed* (Schlosser, 2013 is explicit about this). However, many leading Reasons-Responsive theorists rarely mention consciousness. Indeed, Yaffe claims that "there is no reason to suppose that consciousness is required for reasons-responsiveness" (2012, p.182). (2015a: 931)

Since most reasons-responsive accounts generally ignore or fail to impart a significant role to consciousness, they too can be placed in the third category.

Let me take a moment to briefly discuss Sher and Smith's accounts, since they are representative of the kinds of views that explicitly reject a consciousness requirement on free will. Most accounts of moral responsibility maintain an *epistemic condition* along with a *control condition*—with perhaps some additional conditions added. The former demands that an agent know what they are doing in some important sense, while the latter specifies the kind of control in action needed for moral responsibility. In *Who Knew? Responsibility Without Awareness* (2009), Sher focuses on the epistemic condition and criticizes a popular but inadequate understanding of it (at least according to him). His target is the "searchlight view" which assumes that agents are responsible only for what they are aware of doing or bringing about—i.e., that their responsibility extends only as far as the searchlight of their consciousness. Sher argues that the searchlight view is (a) inconsistent with our attributions of responsibility to a broad range of agents who *should but do not realize* that they are acting wrongly or foolishly, and (b) not independently defensible. Sher defends these criticisms by providing everyday examples of agents who intuitively appear morally responsible but who act for reasons of which they are ignorant or unaware.

As a positive replacement, Sher defends a disjunctive epistemic condition on moral responsibility—one that allows for the possibility of responsibility even in cases of ignorant wrongdoing where one is unaware, so long as the ignorance involved stems from the agent himself in the right way. More specifically, Sher defends the following epistemic condition—which he calls *FEC* for "full epistemic condition" (2009: 143):

> **FEC:** When someone performs an act in a way that satisfies the voluntariness condition, and when he also satisfies any other conditions for responsibility that are independent of the epistemic condition, he is responsible for his act's morally or prudentially relevant feature if, but only if, he either
>
> (1) is consciously aware that the act has the feature (i.e., is wrong or foolish or right or prudent) when he performs it; or else
>
> (2) is unaware that the act is wrong or foolish despite having evidence for its wrongness or foolishness his failure to recognize which
>
> a. falls below some applicable standard, and
> b. is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits; or else
>
> (3) is unaware that the act is right or prudent despite having made enough cognitive contact with the evidence for its rightness or prudence to enable him to perform the act on that basis. (2009: 143)

Sher admits that FEC is "complicated and unlovely" (2009: 144) but maintains that each condition is well motivated. The basic idea behind Sher's view is that the relation between an agent and her failure to recognize the wrongness of what she is doing should be understood in causal terms—i.e., the agent is responsible when, and because, her failure to respond to her reasons for believing that she is acting wrongly has its origins in the same constitutive psychology that generally does render her reasons-responsive.

Angela Smith (2005) likewise argues that we are justified in holding ourselves and others responsible for actions that do not appear to reflect a conscious choice or decision. Her argument is different than Sher's, however, since she attacks the notion that voluntariness (or active control) is a precondition of moral responsibility rather than the epistemic condition. She writes:

> My aim…is to present an alternative to what I have called the volitional view of moral responsibility, one which I think does a better job of capturing the real basis of our responsibility for our own attitudes. Since it is often claimed that the considerations that push us toward the volitional view arise out of our commonsense intuitions about the nature of activity and passivity, however, much of my argument in this article will be devoted to analyzing and rejecting this claim. I will try to show that our commonsense intuitions do not, in fact, favor a volitionalist criterion of responsibility, but a rationalist one. That is to say, I will argue that the kind of activity implied by our moral practices is not the activity of [conscious] choice, but the activity of evaluative judgment. This distinction is important, because it allows us to say that what makes an attitude "ours" in the sense relevant to questions of responsibility and moral assessment is not that we have voluntarily chosen it or what we have voluntary control over it, but that it reflects our own evaluative judgments or appraisals. There are a number of different ways in which our attitudes can be said to reflect our evaluative judgments, but what is important here is

> that a connection to [conscious] choice is not an essential condition of responsibility for these states. (2005: 237)

Smith then proceeds by considering various examples designed to bring out the intuitive plausibility of the "rational relations view," while at the same time casting doubt upon the claim that we ordinarily take conscious choice or voluntary control to be a precondition of legitimate moral assessment.

One set of examples she discusses involves "involuntary" failings—e.g., when I do not notice when my music is too loud, when my advice is unwelcome, or when my assistance might be helpful to others. In such cases, I am unaware of the consequences of my actions, "nevertheless, these failing are commonly taken to be an appropriate basis for moral criticism" (2005: 244). And this is because: "These forms of moral insensitivity provide at least some indication that I do not judge your needs and interests to be important, or at least that I do not take them very seriously" (2005: 244). According to Smith, this helps explain why we normally feel so awful about forgetting important dates, anniversaries, or occasions: "because the normal connection between what occurs to us, on the one hand, and what we care about and judge to be of importance, on the other, we recognize that our failures in these cases can reasonably be taken to reflect a lack of appreciation for the significance of the events in question" (2005: 248).

Smith also discusses responses that are not normally under our immediate conscious control—for example, spontaneous emotions and attitudinal reactions. She writes, for example, "we react with, among other things, envy, admiration, resentment, awe, amusement, regret, and gratitude to the people and events we encounter, and these reactions usually arise without any choice or decision on our part" (2005: 249). Yet, Smith argues, we regularly take these involuntary responses to have a great deal of expressive significance and they often reveal unpleasant and perhaps even painful facts about ourselves—"as when a prejudiced reaction makes us aware of the fact that we have been harboring certain objectionable biases toward others, or a jealous reactions makes us realize that we are distrustful of someone we love" (2005: 249). According to Smith, these examples suggest a direct rational connection between our spontaneous reactions and our underlying evaluative judgments and commitments.

> Attitudes such as contempt, jealousy, and regret seem to be partially constituted by certain kinds of evaluative judgments or appraisals. To feel contempt toward some person, for example, involves the judgment that she has some feature or has behaved in some way which makes her unworthy of one's respect, and to feel regret involves the judgment that something of value has been lost. There seems to be a conceptual connection between having these attitudes and making, or being disposed to make, certain kinds of judgments. This helps to explain why we attach so much significance to these reactions, both in our own case and in our relations to others: unlike brute sensations, which simply assail us, our spontaneous reactions reveal, in a direct and sometimes distressing way, the underlying evaluative commitments shaping our responses to the situations in which we find ourselves. (2005: 250)

According to Smith, examples like these help reveal that, "we attach (moral) significance to a wide variety of attitudes and mental states, many of which do not arise from conscious choice or

decision, and many of which do not seem to fall under our immediate voluntary control" (2005: 250-51).

One potential criticism of Smith's view, however, is that while cases of involuntary failings and spontaneous emotions and reactions may establish that agents are responsible in *some* sense when conscious choice and control is absent, they do not establish that they are responsible in the *basic desert sense*—the sense relevant to the free will debate. It seems intuitive, for instance, to think that agents are responsible in the *attributability sense* in such cases—i.e., we can justifiably attribute these attitudes and failings to the agent him/herself. Attributability responsibility is simply about actions and attitudes being properly attributable to, or reflective of, and agent's self. One may even grant that in Smith's examples agents are also responsible in the *answerability sense*. According to this conception of responsibility, someone is responsible for an action or attitude just in case it is connected to her capacity for evaluative judgment in a way that opens her up, in principle, to demands for justification from others (see Oshana 1997; Scanlon1998; Pereboom 2014). Yet, many philosophers distinguish these conceptions of responsibility from a third kind: *accountability responsibility* or *basic desert moral responsibility* (see Watson 1996; Shoemaker 2011; Eshleman 2014; Caruso 2017). To the extent that basic desert is what the free will debate is primarily concerned with—and I believe it constitutes the core of the debate (Caruso and Morris 2016)—Smith's account might be an answer to the wrong question.[2]

Contrary to these views, Neil Levy (2014), Joshua Shepherd (2012, 2015a), and myself (Caruso 2012, 2015b) have argued that consciousness *is* in fact required for free will and moral responsibility—and accounts like those described above that deny or reject a consciousness condition are untenable, flawed, and perhaps even incoherent. Neil Levy, for example, has argued for something he calls the *consciousness thesis*, which maintains that "consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility" (2014: 1). He contends that since consciousness plays the role of integrating representations, behavior driven by non-conscious representations are inflexible and stereotyped, and only when a representation is conscious "can it interact with the full range of the agent's personal-level propositional attitudes" (2014: vii). This fact entails that consciousness of key features of our actions is a necessary (though not sufficient) condition for moral responsibility since consciousness of the morally significant facts to which we respond is required for these facts to be assessed by and expressive of the agent him/herself.

Furthermore, Levy argues that the two leading accounts of moral responsibility outlined above—*deep self* (or what he calls *evaluative accounts*) and *reasons-responsive* (or *control-based*) accounts—are committed to the truth of the consciousness thesis despite what proponents of these accounts maintain. According to Levy: (a) only actions performed consciously express our

[2] Smith seems to acknowledge this when she writes: "I interpret the fundamental question of responsibility as a question about the conditions of moral attributability, that is to say, the conditions under which something can be attributed to a person in a way that is required in order for it to be a basis for moral appraisal of that person" (2005: 238). *Attributability responsibility*, however, though perhaps a necessary condition, is not sufficient for basic desert responsibility—hence, it cannot alone justify backward-looking praise and blame.

evaluative agency, and that expression of moral attitudes requires consciousness of that attitude; and (b) we possess reasons-responsive control only over actions that we perform consciously, and that control over their moral significance requires consciousness of that moral significance.

In assessing Levy's consciousness thesis, a couple of things are important to keep in mind. First, Levy maintains that the work of Libet and Wegner, which is often cited in conjunction with consciousness and free will, is simply irrelevant to moral responsibility since "it makes no different whether or not consciousness has the powers they contend it lacks" (2014: vii). For instance, Libet's pioneering investigation into the timing of conscious intentions is often interpreted as showing that consciousness is *epiphenomenal*—lacking any causal role in action production (see Section III for discussion). According to Levy, there are two problems with this claim. First, important empirical and interpretative criticisms exist which challenge the purported conclusions of both (see, e.g., Nahmias 2002; Mele 2009; Dennett 1991; Rosenthal 2002; Pereboom and Caruso 2017). Second, and perhaps more importantly, he contends "the consciousness thesis they have been taken to challenge is a different thesis to the one I have in mind" (2014: 15). According to Levy, those who think the work of Libet and Wegner undermine free will and moral responsibility are "wrong in claiming that it is a conceptual truth that free will (understood as the power to act such that we are morally responsible for our actions) requires the ability consciously to initiate action" (2014: 16). Instead, for Levy, what is of true importance is the causal efficacy of deliberation—i.e., "we want it to be the case that our conscious deliberations—our conscious consideration of reasons for and against a particular action—is causally efficacious" (2014: 24). Levy's consciousness thesis therefore demands, not the conscious initiation of action but rather, consciousness of the facts that give our actions their moral significance.

Second, the kind of consciousness Levy has in mind is *not* phenomenal consciousness but rather states with *informational* content. That is, he limits himself to philosophically arguing for the claim that "contents that might plausibly ground moral responsibility are *personally* available for report (under report-conducive conditions) and for driving further behavior, but also occurrent [in the sense of] shaping behavior or cognition" (2014: 31). By *personally available* Levy means the following: "Information is personally available…when the agent is able to effortlessly and easily retrieve it for use in reasoning *and* it is online" (2014: 33). In turn, information is available for easy and effortless recall if "it would be recalled given a large range of ordinary cues; no special prompting (like asking a leading question) is required" (2014: 34). This notion of personal availability is closely akin to what Ned Block (1995) has called *access consciousness*—though the two are not exactly equivalent. Levy prefers not to build into the definition of *personal availability* the fact that the information involved must also be available to a broad variety of consuming systems, whereas Block builds such availability into his definition of access consciousness. But since Levy thinks the empirical question is answered in the affirmative—i.e., personally available information *is* information broadcast to a broad variety of consuming systems in the mind—it turns out that "as a matter of empirical fact the two [notions] are coextensive" (Levy 2014: 35).

We can now say that on Levy's formulation of the consciousness thesis, information of the right kind must be personally available to ground moral responsibility. But what kind of information is the right kind? Here Levy writes, "if the thesis is that agents must be conscious of *all* the mental

states that shape their behavior, no one would ever be responsible for anything" (2014: 36). Rather than demanding consciousness of all relevant mental states, Levy argues that when agents are morally blameworthy or praiseworthy for acting in a certain manner they must be conscious of certain facts which play an especially important role in explaining the *valence* of responsibility. Valence, in turn, is defined in terms of moral significance: "facts that make the action bad play this privileged role in explaining why responsibility is valenced negatively, whereas facts that make the action good play this role in explaining why the responsibility is valenced positively" (2014: 36). Additionally, the morally significant facts that determine the valence need not track the actual state of affairs that pertain, but the facts that the agent *takes* to pertain. The consciousness thesis can now be stated as follows:

> The consciousness thesis is the claim that an agent must be conscious of (what she takes to be) the facts concerning her action that play this important role in explaining its moral valence; these are facts that constitute its moral character. (2014: 37)

According to the consciousness thesis, then, if an action is morally bad the agent must be conscious of (some of) the aspects that make it bad, and conscious of those aspects under appropriate descriptions, in order to be blameworthy for the action.

In my book *Free Will and Consciousness* (Caruso 2012), I also argued for a consciousness thesis, though there I argued for the claim that conscious control and guidance where of utmost importance. That is, I argued that "for an action to be free, consciousness must be involved in *intention* and *goal formation*" (2012: 100). My reasoning was motived by cases of somnambulism and concerns over automaticity and the adaptive unconscious (2012: 100-130) where conscious executive control and guidance are largely absent. More recently, however, I have come to think that Levy's consciousness thesis, or something close to it, is more accurate (see Caruso 2015a, b). This is because, first, I no longer think that the empirical challenges to *conscious will* from neuroscience are all that relevant to the problem of free will (see Pereboom and Caruso 2017). Second, many of the arguments I presented in the book are captured just as well, perhaps better, by Levy's version of the consciousness thesis—including my internal challenge to compatibilism based on recent developments in the psychology, social psychology, and cognitive science (see next section). Finally, Levy's consciousness thesis has the virtue of capturing what I believe is an intuitive component of the epistemic condition on moral responsibility (contra Sher)—i.e., that agents must be aware of important moral features of their choices and actions to be responsible for them.[3] The one remaining difference between us is that I still prefer to understand and explain consciousness in terms of the *Higher-Order Thought* (HOT) theory of consciousness (Caruso 2012, 2005; see also Rosenthal 2005) while Levy favors the *Global Workspace Theory* (Levy 2014; see also Baars 1988, 1997; Dehaene and Naccache 2001; Dehaene, Changeux, and Naccache 2011).

Joshua Shepherd has also argued that consciousness is a necessary condition for free will but his argument is based on taking our folk psychological commitments seriously (2015a). In a series

[3] To be clear, Levy's consciousness thesis is more than just a reformulation of the epistemic condition on moral responsibility. It also highlights important control functions that consciousness provides.

of studies, for example, he has provided evidence that folk views of free will and moral responsibility accord a central place to consciousness (2012, 2015). In one set of studies (2012), he presented participants with vignettes that contrasted the production of behavior by conscious processes with the production of behavior by nonconscious processes. He found that when consciousness was involved in an agent's action production, participants attributed free will and moral responsibility to the agent. When consciousness was not so involved, most participants judged that the agent did not act freely or responsibly. These findings appear to indicate that simply varying the causal impact of consciousness is enough to influence the attribution of free will and moral responsibility. Interestingly, this was true even in a case that varied not only the role of consciousness in action production, but also whether causation was deterministic or indeterministic (2012, experiment 3).

Perhaps more damaging, however, to the views of Arplay, Smith, and Sher, are the findings of Shepherd's second set of studies (2015a). In these experiments, Shepherd contrasted behavior produced by elements of an agent's *deep self*—that is, by elements of an agent's interior life (e.g., motivations, values, and convictions) that the agent clearly endorses—with behavior produced by an agent's conscious mental processes (or "Conscious Self"). Contrary to the intuitions of those deep self theorists who downplay the importance of consciousness, Shepherd's results found that though elements of an agent's deep self do have a minor impact on judgments of free will, ordinary folk consider consciousness more important. For example, in Shepherd's second experiment (2015a), the effect size in the Conscious Self condition was much greater than the Deep Self condition. The results indicate that the impact of consciousness on participant's attributions of free will is both independent of considerations often taken to motivate deep self views, as well as comparatively much stronger (2015a: 938). As a result, Shepherd concludes that "[t]he right interpretation, in my view, is that consciousness is central to folk views of free and responsible action, and that the way in which it is central is not captured by extant Deep Self Views" (2015a: 938)—i.e., views that downplay or neglect the importance of consciousness "contain a significant blind spot" (2015a: 935).

III. If consciousness is necessary for free will, can we ever be free and morally responsible?

Assuming for the moment that consciousness is required for free will, the next question would be: Can the consciousness requirement be satisfied given the threat of shrinking agency and empirical findings in the behavioral, cognitive, and neurosciences? In the literature, two leading empirical threats to the consciousness condition are identifiable. The first maintains that recent findings in neuroscience reveal that unconscious brain activity causally initiates action prior to the conscious awareness of the intention to act and that this indicates *conscious will* is an illusion (e.g., Benjamin Libet, John-Dylan Haynes, Daniel Wegner). The pioneering work in this area was done by Benjamin Libet and his colleagues. In their groundbreaking study on the neuroscience of movement, Libet et al. (1983) investigated the timing of brain processes and compared them to the timing of consciousness will in relation to self-initiated voluntary acts and found that the consciousness intention to move (which they labeled *W*) came 200 milliseconds before the motor act, but 350-400 milliseconds after *readiness potential*—a ramp-like buildup of electrical activity that occurs in the brain and precedes actual movement. Libet and others have interpreted this as showing that the conscious intention or decision to move cannot be the cause of action because it comes too late in the neuropsychological sequence (see Libet 1985, 1999).

According to Libet, since we become aware of an intention to act only after the onset of preparatory brain activity, the conscious intention cannot be the true cause of the action (see also Wegner 2002; Soon et al. 2008; Pockett 2004; Obhi and Haggard 2004; Haggard and Eimer 1999).

Libet's findings, in conjunction with additional findings by John Dylan Haynes (Soon et al. 2008) and others, have led some theorists to conclude that conscious will is an illusion and plays no important causal role in how we act. Haynes and his colleagues, for example, were able to build on Libet's work by using functional magnetic resonance imaging (fMRI) to predict with 60% accuracy whether subjects would press a button with either their right or left hand up to 10 seconds before the subject became aware of having made that choice (Soon et al. 2008). For some, the findings of Libet and Haynes are enough to threaten our conception of ourselves as free and responsible agents since they appear to undermine the causal efficacy of the types of willing required for free will.

Critics, however, maintain that there are several reasons for thinking that these neuroscientific arguments for free will skepticism are unsuccessful. First, critics contend that there is no direct way to tell which conscious phenomena, if any, correspond to which neural events. In particular, in the Libet studies, it is difficult to determine what the readiness potential corresponds to—for example, is it an *intention formation* or *decision*, or is it merely an *urge* of some sort? Al Mele (2009) has argued that the readiness potential (RP) that precedes action by a half-second or more need not be construed as the *cause* of the action. Instead, it may simply mark the beginning of forming an *intention* to act. According to Mele, "it is much more likely that what emerges around -500 ms is a *potential cause* of a proximal intention or decision than a proximal intention or decision itself" (2009: 51). On this interpretation, the RP is more accurately characterized as an "urge" to act or a preparation to act. That is, it is more accurately characterized as the advent of items in what Mele calls the *preproximal-intention group* (or PPG). If Mele is correct, this would leave open the possibility that conscious intentions (or their neural correlates) can still be causes—i.e., if the readiness potential does not correspond to the formation of an intention or decision, but rather an urge, then it remains open that the intention formation or decision is a conscious event.

A second criticism is that almost everyone on the contemporary scene who believes we have free will, whether compatibilist of libertarian, also maintains that freely willed actions are caused by a chain of events that stretch backwards in time indefinitely. At some point in time these events will be such that the agent is not conscious of them. Thus, all free actions are caused, at some point in time, by unconscious events. However, as Eddy Nahmias (2011) correctly points out, the concern for free will raised by Libet's work is that *all* of the relevant causing of action is (typically) nonconscious, and consciousness is not causally efficacious in producing action. Given determinist compatibilism, however, it's not possible to establish this conclusion by showing that nonconscious events that precede conscious choice causally determine action since such compatibilists hold that every case of action will feature such events, and that this is compatible with free will. And given most incompatibilist libertarianisms, it's also impossible to establish this conclusion by showing that there are nonconscious events that render actions more probable than not by a factor of 10% chance (Soon et al., 2008) since almost all such libertarians hold that free will is compatible with such indeterminist causation by unconscious events at some

point in the causal chain (De Caro 2011).

Third, Neil Levy raises a related objection when he criticizes Libet's *impossible demand* (2005) that only consciously initiated actions could be free. Levy correctly argues that this presupposition places a condition upon freedom of action which is in principle impossible to fill for reasons that are entirely conceptual and have nothing to do, per se, with Libet's empirical findings. As Levy notes, "Exercising this kind of control would require that we control our control system, which would simply cause the same problem to arise at a higher-level or initiate an infinite regress of controllings" (2005: 67). If the unconscious initiation of actions is incompatible with control over them, then free will is impossible on conceptual grounds. Thus, Libet's experiments do not constitute a separate, empirical, challenge to our freedom (see Levy 2005).

Several other critics have noted the unusual nature of the Libet-style experimental situation—i.e., one in which a conscious intention to flex at some time in the near future is already in place, and what is tested for is the specific implementation of this general decision. Nahmias (2011), for example, points out that it's often the case—when, for instance, we drive or play sports or cook meals—that we form a conscious intention to perform an action of a general sort, and subsequent specific implementation are not preceded by more specific conscious intentions. But in such cases the general conscious intention is very plausibly playing a key causal role. In Libet-style situations, when the instructions are given, subjects form conscious intentions to flex at some time or other, and if it turns out that the specific implementations of these general intentions are not in fact preceded by specific conscious intentions, this would be just like the kinds of driving and cooking cases Nahmias cites. It seems that these objections cast serious doubts on the potential for neuroscientific studies to undermine the claim that we have the sort of free will at issue.

Finally, as already mentioned above, Levy's version of the consciousness condition—i.e., his *consciousness thesis*—does not require that our voluntary actions be consciously initiated (or caused by conscious intentions or willings) as the neuroscientific argument for free will skepticism assumes. Rather, it only requires that an agent be conscious of some of the facts that give her action its moral significance. Only the latter is required, according to Levy, because the function of *conscious integration* is a necessary (though not sufficient) condition for moral responsibility since consciousness of the morally significant facts to which we respond is required for these facts to be assessed by and expressive of the agent him/herself. If, then, one were to adopt a consciousness condition similar to Levy's—i.e., one that did not require conscious will or the conscious initiation of action—then the concerns of Libet, Wegner, and their followers would become moot. This does not mean, however, that there are no other empirical threats to free will and moral responsibility since it remains an open question to what extent and how often agents satisfy the kind of consciousness condition Levy has in mind. Indeed, the second class of empirical threats to free will is more relevant to accounts like Levy's and others' who suggest that consciousness is essential to free will.

In defending the consciousness thesis, Levy argues that "the integration of information that consciousness provides allows for the flexible, reasons-responsive, online adjustment of behavior." Without such integration, "behaviors are stimulus driven rather than intelligent

responses to situations, and their repertoire of responsiveness to further information is extremely limited" (2014: 39). Consider, for example, cases of *global automatism* (Levy and Bayne 2004). Global automatisms may arise as a consequence of frontal and temporal lobe seizures and epileptic fugue, but perhaps the most familiar example is somnambulism. Take, for instance, the case of Kenneth Parks, the Canadian citizen who on May 24, 1987 rose from the couch where he was watching TV, put on his shoes and jacket, walked to his car, and drove 14 miles to the home of his parents-in-law where he proceeded to strangle his father-in-law into unconsciousness and stab his mother-in-law to death. He was charged with first degree murder but pleaded not guilty, claiming he was sleep-walking and suffering from "non-insane automatism." He had a history of sleepwalking, as did many other members of his family, and the duration of the episode and Parks' fragmented memory were consistent with somnambulism. Additionally, two separate polysomnograms indicated abnormal sleep. At his trial, Parks was found not guilty and the Canadian Supreme Court upheld the acquittal.

While cases like this are rare, they are common enough for the defense of non-insane automatism to have become well established (Fenwick 1990; Schopp 1991; McSherry 1998). Less dramatic, though no less intriguing, are cases involving agents performing other complex actions while, apparently asleep (Levy 2014: 72). Siddiqui et al. (2009), for example, recently described a case of sleep emailing. As Levy describes it: "After the ingestion of zolpidem for treatment of insomnia, the patient arose from bed, walked to the next room and logged onto her email. She then sent three emails in the space of six minutes, inviting friends for dinner and drinks the next day. She had no recall of the episode afterwards" (2014: 72).

These cases illustrate the complexity of the behaviors in which agents may engage in the apparent absence of awareness (Levy 2014: 72). The capacities required for sleep emailing are rather complex: typing a relatively coherent message, entering a subject line, pressing 'send'—all seem to require a high level of cognitive capacity. This all raises the following question: if somnambulism (and other global automatisms) is a disorder of consciousness characterized by a dramatically reduced level of awareness of behavior and surroundings, how can we explain the complex behaviors exhibited by sleep emailers or by Parks? It is here that Levy introduces the notion of *action scripts*:

> Skills the acquisition of which requires the engagement of brainscale distributed networks may be carried out efficiently and in the absence of consciousness, by networks of brain areas that are more localized. The skills that sleep emailing or sleep walking agents exercise are, in the jargon of psychology, overlearned…As a consequence they may be carried out efficiently in the absence of consciousness. Agents who experience disorders of consciousness follow what we might call *action scripts*, which guide their actions, I suggest, where a script is a set of motor representations, typically a chain of such representations, that can be triggered by an appropriate stimulus, and which once triggered runs ballistically to completion. (2014: 74-5)

An example of an action script would be learning to change gears in a manual car: we learn an extended series of movements, each of which serves as the trigger for the next. In acquiring these scripts, we acquire capacities that may allow us to act efficiently in the absence of consciousness.

Levy argues that behaviors driven by action scripts tend to be inflexible and insensitive to vital environmental information. The behaviors of somnambulists, for instance, exhibit some degree of responsiveness to the external environment, but they also lack genuine flexibility of response. To have genuine flexibility of response, or sensitivity to the content of a broad range of cues at most or all times, consciousness is required. With regard to free will and moral responsibility, Levy argues that the functional role of awareness "entails that agents satisfy conditions that are widely plausibly thought to be candidates for necessary conditions of moral responsibility only when they are conscious of facts that give to their actions their moral character" (2014: 87). More specifically, Levy argues that deep self and reasons-responsive accounts are committed to the truth of the consciousness thesis despite what proponents of these account maintain. And this is because (a) only actions performed consciously express our evaluative agency, and that expression of moral attitudes requires consciousness of that attitude; and (b) we possess responsibility-level control only over actions that we perform consciously, and that control over their moral significance requires consciousness of that moral significance.

Consider again the Kenneth Parks case. Assuming that Parks was in a state of global automatism on the night of May 24, 1987, he acted without consciousness of a range of facts, each of which gives to his actions moral significance—"he is not conscious *that he is stabbing an innocent person*; he is not conscious *that she is begging him to stop*, and so on" (2014: 89). These facts, argues Levy, "entail that his actions do not express his evaluative agency or indeed any morally condemnable attitude" (2014: 89). Because Park is not conscious of the facts that give to his actions their moral significance, these facts are not globally broadcast—and because these facts are not globally broadcast, "they do not interact with the broad range of the attitudes constitutive of his evaluative agency" (2014: 89). This means that they do not interact with his *personal-level* concerns, beliefs, commitments, or goals. Because of this, Levy maintains that Parks' behavior is "not plausibly regarded as an expression of his evaluative agency"—agency caused or constituted by his personal-level attitudes (2014: 90).

Now, it's perhaps easy to see why agents who lack creature consciousness, or are in a very degraded global state of consciousness, are typically excused moral responsibility for their behaviors, but what about more common everyday examples where agents *are* creature conscious but are not conscious of a fact that gives an action its moral significance? Consider, for instance, an example drawn from the experimental literature on implicit bias. Uhlmann and Cohen (2005) asked subjects to rate the suitability of two candidates for police chief, one male and one female. One candidate was presented as "streetwise" but lacking in formal education, while the other one had the opposite profile. Uhlmann and Cohen varied the sex of the candidates across conditions, so that some subjects got a male streetwise candidate and a female well-educated candidate, while other subjects got the reverse. What they found was that in both conditions subjects considered the male candidate significantly better qualified than the female, with subjects shifting their justification for their choice. That is, they rated being "streetwise" or being highly educated as a significantly more important qualification for the job when the male applicant possessed these qualifications than when the female possessed them. These results indicate "a preference for a male police chief was driving subjects' views about which characteristics are needed for the job, and not the other way around" (Levy 2014: 94).

Is this kind of implicit sexism reflective of an agent's *deep self* such that he should be held morally responsible for behaviors stemming from it? Levy contend that, "though we might want to say that the decision was a sexist one, its sexism was neither an expression of evaluative agency nor does the attitude that causes it have the right kind of content to serve as grounds on the basis of which the agent can be held (directly) morally responsible" (2014: 94). Let us suppose for the moment that the agent does not consciously endorse sexism in hiring decisions—i.e., that had the agent been conscious that the choice had a sexist content he would have revised or abandoned it. Under this scenario, the agent was not conscious of the facts that give his choice its moral significance. Rather, Levy argues, "they were conscious of a confabulated criterion, which was itself plausible (it is easy to think of plausible reasons why being streetwise is essential for being police chief; equally, it is easy to think of plausible reasons why being highly educated might be a more relevant qualification)" (2014: 95). Since it was this confabulated criterion that was globally broadcast (in the parlance of Levy's preferred global workspace theory of consciousness), and which was therefore assessed in the light of the subjects' beliefs, values, and other attitudes, the agent was unable to evaluate and assess the implicit sexism against his personal-level attitudes. It is for this reason that Levy concludes that the implicit bias is "not plausibly taken to be an expression of [the agent's] evaluative agency, their deliberative and evaluative perspective on the world" (2014: 95).

Levy makes similar arguments against reasons-responsive accounts of moral responsibility. He argues that in both the case of global automatism and implicit bias, reasons-responsive control requires consciousness. This is because (a) reasons-responsiveness requires creature consciousness, and (b) the agent must be conscious of the moral significance of their actions in order to exercise responsibility-level control over it. In the Kenneth Parks case, for example, his behavior may be *weakly* responsive to reasons—i.e., "there is some scenario in which the mechanisms that cause behavior would be receptive and reactive to a reason to do otherwise" (Levy 2014: 112)—but weak reasons-responsiveness is not sufficient for guidance control. Fischer and Ravizza (1998), for instance, require *moderate* reasons-responsiveness. On their account, a mechanism is moderately reasons-responsive when it is regularly receptive to reasons. That is, the mechanism must be responsive to reasons, including moral reasons, in an understandable pattern. According to Levy, this condition entails that "agents like Parks do not exercise guidance control over their behavior, because the mechanism upon which they act (the action script) is not regularly receptive to reasons" (2014: 113).

The case of implicit bias also fails to meet the requirement of moderate reasons-responsiveness. Considering again the Uhlmann and Cohen example of implicit sexism, Levy writes:

> These subjects were, of course, conscious agents, but they were…not conscious of the implicit attitudes that biased their information processing, thereby producing their confabulated criteria for job suitability. This implicit attitude imparted to their decision its moral significant content: its sexism. But because the subjects were conscious neither of the attitude nor its effect on their decision, they could not detect conflicts between either their attitudes or their decisions, on the one hand, and their personal-level attitudes, on the other. What was globally broadcast, and therefore assessed for consistency and conflict, was the confabulated criterion; the attitude that caused the confabulation was neither broadcast nor assessed. (Levy 2014: 115)

Because these agents were conscious neither of the implicit attitude that caused the confabulation, nor of the moral significance of the decision they made, Levy maintains that they could not exercise guidance control over either. Hence, if moral responsibility requires guidance control, then once again agents like those discussed above are excused moral responsibility.

Levy's defense of the consciousness condition and his assessment of the two leading accounts of moral responsibility entail that people are less responsible than we might think. But how much less? In the final section of his book he address the concerns of theorists like myself (Caruso 2012) who worry that the ubiquity and power of non-conscious processes either rule out moral responsibility completely or severely limit the instances where agents are justifiably blameworthy and praiseworthy for their actions. There he maintains that adopting the consciousness thesis need not entail skepticism of free will and basic desert moral responsibility since the consciousness condition can be (and presumably often is) met. His argument draws on an important distinction between cases of global automatism and implicit bias, on the one hand, and cases drawn from the *situationist* literature on the other. Levy maintains that in the former cases (global automatism and implicit bias), agents are excused moral responsibility since they either lack creature consciousness or they are creature conscious but fail to be conscious *of* some fact or reason which nevertheless plays an important role in shaping their behavior. In situational cases, however, Levy maintains that agents *are* morally responsible despite the fact that their actions are driven by non-conscious situational factors, since the moral significance of their actions remains consciously available to them and globally broadcast (Levy 2014: 132).

In a response to Levy, I have argued that if consciousness is necessary for free will and moral responsibility, then people are *significantly* less responsible than we think (Caruso 2015b). My argument distinguishes between three types of cases defined as follows:

(1) Type-1 Cases: These are cases of *global automatism*, where agents either lack creature consciousness altogether or are in a very degraded state of consciousness. These cases are dramatic, puzzling, and relatively rare. Examples include cases of somnambulism such as the Kenneth Parks case.

(2) Type-2 Cases: Far more common are cases of agents who are normally conscious (creature conscious), but fail to be conscious *of* some fact or reason which nevertheless plays a significant role in shaping their behavior. Examples include favoring a male candidate over a female candidate because of *implicit sexism* (Uhlmann and Cohen 2005) and other examples of implicit bias.

(3) Type-3 Cases: Perhaps even more common still are cases where agents are conscious of facts that shape their behavior, but conscious nether of *how*, nor even *that*, those facts shape their behavior. Examples of type-3 cases can be found in the *situationist literature*—for example, an agent may be conscious that they previously held a hot cup of coffee, but not conscious *that* (or *how*) the cup of coffee affected their judgment of others (Williams and Bargh 2008).

Levy maintains that agents are excused moral responsibility in type-1 and type-2 cases but *not* in type-3 cases, while I have argued that type-3 cases can *also* fail to satisfy the consciousness thesis (see Caruso 2015b). By extending the realm of morally excusable cases to type-3 cases, I do not mean to suggest that *all* moral responsibility would be ruled out (at least not for reasons having to do with consciousness).[4] It remains an open empirical question the extent to which our choices and actions are affected in type-3 ways. That said, there is no doubt that adopting such a view would severely limit the cases were agents could be held morally responsible since type-3 cases are common and unexceptional.

My argument for extending skepticism to type-3 cases can be found in Caruso (2015b). There I consider several examples drawn from the situationist literature—each one representing a case in which agents are conscious of facts that shape their behavior, but conscious neither of *how*, nor even *that*, those facts shape their behavior—and argue that we should excuse moral responsibility in these cases for the same reason Levy provides in type-1 and type-2 cases. For sake of space, however, I will here only consider one example. While some of the situationist priming literature has recently been thrown into doubt due to the so-called "replication crisis," I will take the following example as representative of the broader literature and leave it as an open empirically question how powerful such effects are (for an excellent survey of the literature, see Doris 2002). A recent meta-analysis of the priming literature, however, has found a behavioral priming effect, "which was robust across methodological procedures and only minimally biased by the publication of positive (vs. negative) results" (Weingarten et al. 2016).

Experiments carried out by Bargh, Chen and Burrows (1996) found that when trait constructs were non-consciously activated during an unrelated task, what is known as priming, participants were subsequently more likely to act in line with the content of the primed trait construct. In one experiment, participants were primed on the traits of either rudeness or politeness (or neither) using a scramble-sentence test in which they were told to form grammatical sentences out of short lists of words. Participants were exposed to words related to either rudeness (e.g., rude, impolite, obnoxious), politeness (e.g., respect, consideration, polite), or neither. Participants were told after completing the test that they were to go tell the experimenter they were done. When they attempted to do so, however, the experimenter was engaged in a staged conversation. Bargh and his colleagues wanted to see if participants would interrupt. They found that among those primed for "rudeness" 67% interrupted, among those primed for "politeness" only 16% interrupted, and for the control group 38% interrupted. In addition, during an extensive post-

[4] I should mention that I am a global skeptic about free will and moral responsibility for reasons having nothing to do with consciousness (see Caruso 2012, 2013; Pereboom and Caruso, forthcoming). The empirical case being made here, however, is independent of my arguments for global skepticism and remain in place even if they are rejected. As I see it, there are *external* and *internal* challenges to free will. External challenges target the justification of the whole moral responsibility system—familiar examples include the arguments for hard determinism and hard incompatibilism. Internal challenges, on the other hand, play according to the rules of, say, reasons-responsive or deep self accounts of free will and moral responsibility but challenge them from the inside. I am here only concerned with the latter.

experiment debriefing, none of the participants showed any awareness or suspicion of the possible influence of the scramble-sentence test on their interrupting behavior.[5]

In this situation, the 67% who behave rudely by interrupting a conversation do so because of situational factors the influence of which he/she is not conscious. Should the agent be excused moral responsibility in such situations? I contend that if we apply the same considerations we did in type-2 cases, we should answer in the affirmative. First of all, it is *prima facie* plausible to think that an agent in this situation fails to be conscious of the facts that give his action its moral significance. It's reasonable to think that rather than being conscious *that he is acting rudely* (under this or a similar description), the agent is instead conscious of some confabulated reason for his behavior. Like the implicit sexism case discussed above, this would mean that rather than being conscious of the primed trait construct for rudeness, the agent is conscious of a confabulated reason for his behavior which itself seems plausible. In turn, it would be this confabulated reason that is globally broadcast and assessed against the agent's beliefs, values, and other attitudes. One might even imagine that *if* the agent were aware of the influence of the rudeness prime, he would disapprove of it. Since the agent is unaware of the influence the primed trait construct for rudeness is having on his behavior, he is unable to evaluate and assess it against his personal-level attitudes. Hence, we should conclude (for reasons similar to Levy's) that it is not a reflection of his evaluative agency.

Furthermore, since the agent is conscious neither of the situational factor that caused the confabulation, nor of the moral significance of the behavior, he is unable to exercise guidance control over either. This would be for the very same reason Levy explains when discussing type-2 cases. Guidance control requires moderate reasons-responsiveness, and moderate reasons-responsiveness requires regular receptivity to reasons, including moral reasons. But as Levy himself notes, "[i]nsofar as our behavior is shaped by factors of which we are unaware, we cannot respond to these facts, nor to the conflict or consistency between these facts and other reasons" (2014: 115). We exercise guidance control over those facts of which we are conscious, assessing them as reasons for us, but in this scenario the contents that came up for assessment were confabulated, and the contents that caused the confabulation could not be recognized as reasons. (For Levy's reply, see his 2015).

III. The Introspective Argument for Free Will

Despite the arguments in the previous section 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. Libertarians, for example, 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 this common form of libertarian reasoning: "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

[5] I should note that the situationist literature also includes cases where word priming is not used and participants are instead influenced by mundane physical objects—for example, the mere presence of a briefcase sitting on a desk (see Kay et al. 2004).

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). I have elsewhere called this the *introspective argument* for free will and provided arguments against it (Caruso 2012, Ch.5). Given the limited space remaining, however, I will here only cite a few instances of the introspective argument in the libertarian literature and briefly point to some potential problems with it.

The introspective argument essentially maintains that, upon introspection, we do not *seem* to be causally determined—indeed, 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 not 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 do 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. (1995: 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 deliberation, 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)

This kind of introspective argument is extremely important for libertarians since nearly all assign introspective evidence some role, "for it is our feeling of metaphysically open branching paths that is the raise d'etre of libertarian freedom" (Ross 2006: 135).

There are, however, good reasons for questioning the introspective argument. First, the argument only works if we assume the introspective 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? Even Richard

Taylor acknowledges that the introspective data might simply be an illusion. He writes, for instance:

> One could hard affirm such a theory of agency with complete comfort, however, and not wholly without embarrassment, for the conception of agents and their powers which is involved in it is strange indeed, if not positively mysterious. In fact, one can hardly be blamed here for simply denying our data outright, rather than embracing this theory to which they do most certainly point. Our data…rest upon nothing more than fairly common consent. These data might simply be illusions. (1992: 53)

Moreover, some philosophers have argued that there *is* in fact good reason for concluding that the phenomenology of free agency is illusory or at least seriously misleading. In *Free Will and Consciousness* (Caruso 2012), I identified four phenomenological features of experience responsible for generating our (mistaken) feeling of freedom:

(1) The *apparent* transparency and infallibility of consciousness—i.e., the feeling that we have immediate and direct access to our mental states and processes (at least those relevant to our choices and reasons for action).

(2) The *feeling* that our intentional states arise spontaneously and are causally undetermined or determined only by the agent—in contrast, say, to our sensory states, which are experienced as caused by states of the world.

(3) The *feeling* of *conscious will*—i.e., the feeling that we consciously cause or initiate behavior directly through our conscious intentions, decisions, and willings.

(4) Our *sense* of a *unified self* who is the willful author of behavior. (This includes both a *sense of ownership* and a *sense of authorship*.)

I argued that each of these phenomenological components plays an important role in generating our sense of libertarian freedom but that there are good philosophical and empirical reasons for concluding that each is illusory in important ways—i.e., ways that would undermine the introspective argument.[6]

Consider, for example, the apparent transparency and infallibility of consciousness. From the first-person perspective, we feel as though we have direct access to all the relevant causal factors and processes underlying our own decision-making. Descartes even made transparency and infallibility a key component of his theory of mind, along with his dualism of course. Peter Carruthers (2008) has gone so far as to suggest that a belief in the transparency of the mind is both species-universal and innate. And Benjamin Kozuch and Shaun Nichols (2011) have shown that ordinary folk share this Cartesian belief in introspective transparency—at least for a certain

[6] Whereas I argue that the phenomenology of free agency is illusory, I should note that other philosophers like Tim Bayne and Neil Levy (2006) deny that the phenomenology has the content the libertarian claims. See also Deery (2015). This view, however, would also be a threat to the introspective argument for libertarian freedom.

domain of mental events, namely those related to decision-formation. Yet despite the phenomenological appeal of the belief in introspective transparency, we now know that consciousness is neither transparent nor infallible. Abundant evidence now exists that unconscious states and processes commonly influence our judgments, attitudes, and decisions in ways we are remain completely unaware. Cognitive science indicates that much of our mental lives are not available to introspection (e.g., Nisbett and Wilson 1977; Gopnick 1993; Wegner 2002; Wilson 2002). We also know, from a number of different domains of investigation, that individuals often confabulate reasons for judging, deciding, and acting as they do—examples can be found in social psychology (Nisbett and Wilson 1977), neuroscience (Gazzaniga 1985; Damasio 1994), and the social intuitionist literature (Haidt 2001).

I have argued (Caruso 2012) that since people hold this belief in introspective transparency (regardless of whether or not it is true), they are lead through the phenomenology described above to infer that they are undetermined (see also Nichols 2004). 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. I argued that the belief in the introspective transparency and infallibility of consciousness, coupled with a failure to introspect any deterministic processes underlying our own decision making, contributes to our sense of free will. From the first-person point of view, we feel as though consciousness is immediate, direct, and transparent. The *apparent* transparency 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. But 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. (See Caruso (2012) for additional arguments against the other three phenomenological components.)

A second reason to be dubious of the introspective argument is that while libertarians put a great deal of emphasis on consciousness when it comes to introspecting our own freedom, they often overlook the importance of consciousness when it comes to explaining its role in *producing* free actions. Of course, phenomenology is not enough to establish a metaphysical conclusion. To conclude that we actually have libertarian free will, we would need independent reason for thinking that our experience of free agency corresponds to actual conscious powers and abilities. O'Connor, for example, seems to be aware of this problem 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 automata could conceivably satisfy the conditions set by these accounts—something that is very counterintuitive. (2000: 122)

To his credit, O'Connor seems to recognize that if libertarian accounts of freedom are to be successful, they *must* show that this is not the case and that at least one of the functions of consciousness is to facilitate libertarian free will. For example, with regard to the agent-causal power of self-determination, O'Connor writes:

> 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)

Unfortunately, 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. This general failure can be found throughout the libertarian agent-causal literature.

Libertarian event-causal theorists occasionally say more, but I believe they too fail to give a comprehensive and convincing account. Both John Searle (2000, 2001a, b) and David Hodgson (2005, 2012), for example, advocate *indeterminist* accounts of free will while at the same time rejecting the metaphysical commitments of libertarian agent-causation. They both maintain that consciousness is physically realized at the neurobiological level and advocate naturalist accounts of the mind. Yet they also maintain that there is true (not just psychological) indeterminism involved in cases of rational, conscious decision-making. John Searle's *indeterminist* defense of free will, for example, is predicated on an account of what he calls *volitional consciousness*. According to Searle, consciousness is essential to rational, voluntary action. He boldly proclaims: "We are talking about conscious processes. The problem of freedom of the will is essentially a problem about a certain aspect of consciousness" (2000: 9). Searle argues that to make sense of our standard explanations of human behavior, explanations that appeal to reasons, we have to postulate "an entity which is conscious, capable of rational reflection on reasons, capable of forming decisions, and capable of agency, that is, capable of initiating actions" (2000: 10). Searle maintains that the problem of free will stems from volitional consciousness—our consciousness of the apparent gap between determining reasons and choices. We experience the gap when we consider the following: (1) our reasons and the decision we make, (2) our decision and action that ensues, (3) our action and its continuation to completion (2007: 42). Searle believes that, if we are to act freely then our experience of the gap cannot be illusory: it must be the case that the causation at play is non-deterministic.

Searle attempts to make sense of these requirements by arguing that consciousness is a system feature (see 2000, 2001a) and that the whole system moves at once, but not on the basis of causally sufficient conditions. He writes:

> What we have to suppose, if the whole system moves forward toward the decision making, and toward the implementation of the decision in actual actions; that the conscious rationality at the top level is realized all the way down, and that means that the whole system moves in a way that is causal, but not based on causally sufficient conditions. (2000: 16)

According to Searle, this account is only intelligible "if we postulate a conscious rational agent, capable of reflecting on its own reasons and then acting on the basis of those reasons" (2000: 16). That is, this "postulation amounts to a postulation of a self. So we can make sense of rational, free conscious actions, only if we postulate a conscious self" (2000: 16). For Searle, this means that you cannot account for the rational self just in terms of a Humean bundle of perception. For Searle, the *self* is a primitive feature of the system that cannot be reduced to independent components of the system or explained in different terms.

David Hodgson (2005, 2012) presents a similar defense of free will. As the title of his book states his thesis: *Rationality + Consciousness = Free Will* (2012). According to Hodgson, "in significant choices we are consciously aware of experiences, thoughts (including thoughts in which we attend to *beliefs*), and/or feelings, that provide *reasons*…for one or more available alternatives" (2005: 6). On Hodgson's account, a free action is determined by the conscious subject him/herself and not by external or unconscious factors. He puts forth the following *consciousness requirement*, which he maintains is a requirement for any intelligible account of indeterministic free will: "[T]he transition from a pre-choice state (where there are *open alternatives* to choose from) to a single post-choice state is a conscious process, involving the interdependent existence of a subject and contents of consciousness." For Hodgson, this associates the exercise of free will with consciousness and "adopts a view of consciousness as involving the interdependent existence of a self or subject and contents of consciousness" (2005: 4). In the conscious transition process from pre- to post-choice, Hodgson maintains, the subject grasps the availability of alternatives and knows-how to select one of them. This, essentially, is where free will gets exercised. For Hodgson, it is essential to an account of free will that subjects be considered as capable of being *active*, and that this activity be reflected in the contents of consciousness. He writes: "Again, this is intelligible and plausible: indeed, it is widely accepted that voluntary behavior is active conscious behavior" (2005: 5).

There are, however, several important challenges confronting such accounts. First, Searle and Hodgson's understanding of the *self* is hard to reconcile with our current understanding of the mind, in particular with what we have learned from cognitive neuroscience about reason and decision-making. While it is perhaps true that we *experience* the self as they describe, our sense of a unified self, capable of acting on conscious reasons, may simply be an illusion (see, e.g., Dennett 1991; Caruso 2012). Furthermore, work by Daniel Kahneman (2011) and Jonathan Haidt (2001, 2012) has shown that much of what we take to be "unbiased conscious deliberation" is at best rationalization. Second, Searle's claim that the *system itself* is indeterminist makes sense only if you think a quantum mechanical account of consciousness (or the system as a whole) can be given. He writes, "the lack of causally sufficient conditions at the psychological level goes all the way down. That will seem less puzzling to us if we reflect that our urge to stop at the level of the neurons is simply a matter of prejudice. If we keep on going down to the quantum mechanical level, then it may seem less surprising that we have an absence of causally sufficient conditions" (2000: 17). This appeal to quantum mechanics to account for conscious, rational behavior is problematic for three reasons.

First, it is an empirically open question whether quantum indeterminacies can play the role needed on this account. Max Tegmark (1999), for instance, has argued that in systems as massive, hot, and wet as neurons of the brain, any quantum entanglements and indeterminacies

would be eliminated within times far shorter than those necessary for conscious experience. Tagmark presents calculations to suggest that any macroscopic quantum entanglements in the brain would be destroyed in times of the order of 10(-13) to 10(-20) seconds; far short of what would be required for consciousness. The time scale in typical experiments about consciousness—attention, decision, short-term recall—are generally on the scale of 10(-3). Furthermore, *even if* quantum indeterminacies could occur at the level needed to affect consciousness and rationality, they would also need to exist at precisely the right *temporal* moment—for Searle and Hodgson this corresponds to the gap between determining reasons and choice. As Searle explains, this means "the brain is such that the conscious self is able to make and carry out decisions in the gap, where neither decision nor action is determined in advance, by causally sufficient conditions, yet both are rationally explained by the reasons the agent is acting on" (2007: 73). These are not inconsequential empirical claims—in fact, Searle acknowledges that there is currently no proof for them.

Second, Searle and Hodgson's appeal to quantum mechanics and the way they motivate it comes off as desperate. When Searle, for instance, asks himself, "How could the behavior of the conscious brain be indeterminist? How exactly would the neurobiology work on such an hypothesis?" He candidly answers, "I do not know the answer to that question" (2000: 17). Positing one mystery to account for another mystery, however, is thoroughly unsatisfying and unconvincing.

Lastly, it's unclear that appealing to quantum indeterminacy in this way is capable of preserving free will in any meaningful way. There is a long-standing and very powerful objection against such theories. The *luck objection* (or *disappearing agent objection*) maintains that if our actions are the result of indeterminate events, then they become matters of luck or chance in a way that undermines our free will (see, e.g., Mele 1999; Haji 1999; Pereboom 2001, 2014; Levy 2011; Caruso 2015). As Derk Pereboom described the concern:

> Consider a decision that occurs in a context in which the agent's moral motivations favor that decision, and her prudential motivations favor her refraining from making it, and the strength of these motivations are in equipoise. On an event-causal libertarian picture, the relevant causal conditions antecedent to the decision, i.e., the occurrence of certain agent-involving events, do not settle whether the decision will occur, but only render the occurrence of the decision about 50% probable. In fact, because no occurrence of antecedent events settles whether the decision will occur, and only antecedent events are causally relevant, *nothing* settles whether the decision will occur. Thus it can't be that the agent or anything about the agent settles whether the decision will occur, and she therefore will lack the control required for basic desert moral responsibility for it. (2014: 32)

The core objection is that because event-causal libertarian agents will not have the power to *settle* whether the decision will occur, they cannot have the role in action basic desert moral responsibility demands. Without smuggling back in mysterious agent-causal powers, what does it mean to say that the agent "selects" one set of reasons (as her motivation for action) over another? Presumably this "selection" is not within the active control of the agent since it is the result of indeterminate events that the agent has *no ultimate control over.*

IV. Conclusion

In this survey I have provided a rough taxonomy of views regarding the relationship between consciousness, free will, and moral responsibility. We have seen that there are four broad categories of views, which divide on how they answer the following two questions: (1) Is consciousness necessary for free will? And if so, (2) can the consciousness requirement be satisfied given the threat of shrinking agency and recent developments in the behavioral, cognitive, and neurosciences? With regard to the first question, we find two general sets of views—those that reject and those that accept a consciousness condition on free will. The first group explicitly denies that consciousness is needed for agents to be free and moral responsible but disagree on the reasons why. The second group argues that consciousness *is* required, but then divides further over whether and to what extent the consciousness requirement can be satisfied. I leave it to the reader to decide the merits of each of these accounts. In the end I leave off where I began, with questions: Is consciousness necessary for free will and moral responsibility? If so, what role or function must it play? Are agents morally responsible for actions and behaviors that are carried out automatically or without conscious control or guidance? And are they morally responsible for actions, judgments, and attitudes that are the result of implicit biases or situational features of their surroundings of which they are unaware? These questions need more attention in the literature, since clarifying the relationship between consciousness and free ill is imperative if one wants to evaluate the various arguments for and against free will.

Acknowledgements

I would like to thank Neil Levy, Bruce Waller, Farah Focquaert, Joshua Shepherd, and Eddy Nahmias for helpful comments on an earlier draft of this chapter.

References

Arplay, N. 2002. *Unprincipled Virtues: An inquiry into moral agency*. New York: Oxford University Press.

Arplay, N. and T. Schroeder. 1999. Prais, blame and the whole self. *Philosophical Studies* 93 (2): 161-199.

Baar, B. 1988. *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

Baar, B. 1997. *In the theater of consciousness*. New York: Oxford University Press.

Bargh, J.A. 1997. The automaticity of everyday life. In *The automaticity of everyday life: Advances in social cognition,* Vol.10, ed. Robert S. Wyer, Jr., 1-61. Mahwah, NJ: Erlbaum.

Bargh, J.A. 2008. Free will is un-natural. In *Are we free? Psychology and free will*, ed. John Baer, James C. Kaufman, and Roy F. Baumeister, 128-54. New York: Oxford University Press.

Bargh, J.A., and T.L. Chartrand. 1999. The unbearable automaticity of being. *American Psychologist* 54 (7): 462-79.

Bargh, J., M. Chen, and L. Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71: 230-244.

Bargh, J.A., and M.J. Ferguson. 2000. Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin* 126 (6): 925-45.

Bayne, T. and N. Levy. 2006. The feeling of doing: Deconstructing the phenomenology of agency. In eds. N. Sebanz and W. Prinz, *Disorders of volition*, 49-68. Cambridge, MA: MIT Press.

Baumeister, R.F. 2008. Free will in scientific psychology. *Perspectives of Psychological Science* 3 (1): 14-19.

Baumeister, R.F., A. Mele, and K.D. Vohs. Eds. 2010. *Free will and consciousness: How might they work?* New York: Oxford University Press.
Block, N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18: 227-87.

Campbell, C. A. 1957. *Of selfhood and godhood*. London: Allen and Unwin.

Carruthers, P. 2008. Cartesian epistemology: Is the theory of the self-transparent mind innate? *Journal of Consciousness Studies* 15 (4): 28-53.

Caruso, G.D. 2012. *Free will and consciousness: A determinist account of the illusion of free will*. Lanham, MD: Lexington Books.

Caruso, G.D. Ed. 2013. *Exploring the illusion of free will and moral responsibility*. Lanham, MD: Lexington Books.

Caruso, G.D. 2015a. Précis of Neil Levy's *Consciousness and Moral Responsibility*. *Journal of Consciousness Studies* 22 (7-8): 7-15.

Caruso, G.D. 2015b. If consciousness is necessary for moral responsibility, then people are less responsible than we think. *Journal of Consciousness Studies* 22 (7-8): 49-60.

Caruso, G.D. 2015c. Kane is Not Able: A Reply to Vicens' "Self-Forming Actions and Conflicts of Intention." *Southwest Philosophy Review* 31, 2.

Caruso, G.D. 2017. Free will skepticism and the question of creativity: Creativity, desert, and self-creation. *Ergo* 3 (23).

Caruso, G.D. and O. Flanagan. Eds. 2017. *Neuroexistentialism: Meaning, morals, and purpose in the age of neuroscience.* New York: Oxford University Press.

Caruso, G.D. and S.G. Morris. 2016. Compatibilism and Retributive Desert Moral Responsibility: On What is of Central Philosophical and Practical Importance. *Erkenntnis*. DOI 10.1007/s10670-016-9846-2.

Damasio, A. 1994. *Descartes' error: Emotion, reason, and the human brain.* New York: Avon Books.

Davies, P.S. 2009. *Subjects of the world: Darwin's rhetoric and the study of agency in nature*. Chicago: University of Chicago Press.

De Caro, M. 2011. Is emergence refuted by the neurosciences? The case of free will. In A. Corradini and T. O'Connor (Eds.), *Emergence in science and philosophy*, pp.190-21. London: Routledge.

Deery, O. 2015. The fall from Eden: Why libertarianism isn't justified by experience. *Australasian Journal of Philosophy* 93 (2): 319-334.

Dehaene, S. and L. Naccache. 2001. Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* 79: 1-37.

Dehaene, S., J.P. Changeux, and L. Naccache. 2011. The global neuronal workspace model of conscious access: From neuronal architecture to clinical applications. In S. Dehaene and Y. Christen, eds., *Characterizing consciousness: From cognition to the clinic?* Berlin: Springer-Verlag.

Dennett, D.C. 1991. *Consciousness explained*. London: Penguin Books.

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

Double, R. 1991. *The non-reality of free will*. Oxford: Oxford University Press.

Eshleman, A. 2014. Moral responsibility. *Stanford Encyclopedia of Philosophy*.

Faraci, D. and D. Shoemaker. 2010. Insanity, deep selves, and moral responsibility: The case of JoJo. *Review of Philosophy and Psychology* 1 (3): 319-332.

Feinberg, J. 1970. Justice and personal desert. In his *Doing and deserving*. Princeton: Princeton University Press.

Fenwick, P. 1990. Automatism, medicine and the law. *Psychological Medicine Monograph* 17: 1-27.

Fischer, J.M. 2007. Compatibilism. In *Four Views on Free Will*, eds. J. Fischer, R. Kane, D. Pereboom and M. Vargas, 44-84. Hoboken, NJ: Wiley-Blackwell Publishing.

Fischer, J.M. and M. Ravizza. 1998. *Responsibility and control: A theory of moral responsibility.* Cambridge: Cambridge University Press.

Gazzaniga, M. 1985. *The social brain*. New York: Free Press.

Gopnik, A. 1993. How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Science* 16: 1-14.

Haggard, P. M. Eimer. 1999. On the relation between brain potentials and the awareness of voluntary movement. *Experimental Brain Research* 126 (1): 128-33.

Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108:814-34.

Haidt, J. 2012. *The righteous mind: Why good people are divided by politics and religion.* New York: Pantheon.

Haji, I. 1999. Indeterminism and Frankfurt-type examples. *Philosophical Explorations* 1: 42-58.

Hodgson, D. 2002. Quantum physics, consciousness, and free will. In *The Oxford handbook of free will*, ed. Robert Kane, 85-110. New York: Oxford University Press.

Hodgson, D. 2005. A plain person's free will. *Journal of Consciousness Studies* 12 (1): 1-19.

Hodgson, D. 2008. A role for consciousness. *Philosophy Now* 65: 22-24.

Hodgson, D. 2012. *Rationality + consciousness = free will.* New York: Oxford University Press.

Kahneman, D. 2011. *Thinking fast and slow.* New York: Farrar, Straus, and Giroux.

Kane, R. 1996. *The significance of free will*. New York: Oxford University Press.

Kay, A.C., S.C. Wheeler, J.A. Bargh, and L. Ross. 2004. Material priming: The influence of mundane physical objects on situational construal and competitive behavioral choice. *Organisational Behaviour and Human Decision Processes* 95: 83-96.

King, M. and P. Carruthers. 2012. Moral responsibility and consciousness. *Journal of Moral Philosophy* 9:200-228.

Kozuch, B. and S. Nichols. 2011. Awareness of unawareness: Folk psychology and introspective transparency. *Journal of Consciousness Studies* 18, 11-12: 135-60.

Mele, A. 1999. Ultimate responsibility and dumb luck. *Social Philosophy and Policy* 16: 274-293.

Levy, N. 2005. Libet's impossible demand. *Journal of consciousness studies* 12(12): 67-76.

Levy, N. 2011. *Hard luck: How luck undermines free will and moral responsibility*. New York: Oxford University Press.

Levy, N. 2014. *Consciousness and moral responsibility*. New York: Oxford University Press.

Levy, N. 2015. Defending the consciousness thesis: A response to Robichaud, Sripada and Caruso. *Journal of Consciousness Studies* 22 (7-8): 61-76.

Levy, N. and T. Bayne. 2004. A will of one's own: Consciousness, control and character. *International Journal of Law and Psychiatry* 27: 459-470.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Science* 8:529-66.

Libet, B. 1999. Do we have free will? *Journal of Consciousness Studies* 6 (8-9): 47-57. Reprinted in *The Oxford handbook of free will*, ed. Robert Kane, 551-64. New York: Oxford University Press, 2002.

Libet, B., C.A. Gleason, E.W. Wright, and D. K. Pearl. 1983. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain* 106: 623-42.

McSherry, B. 1998. Getting away with murder: Dissociative states and criminal responsibility. *International Journal of Law and Psychiatry* 21: 163-176.

Mele, A. 1999. Ultimate responsibility and dumb luck. *Social Philosophy and Policy* 16: 274-93.

Mele, A. 2009. *Effective intentions*. New York: Oxford University Press.

Nadelhoffer, T. 2011. The threat of shrinking agency and free will disillusionism. In *Conscious will and responsibility: A tribute to Benjamin Libet*, ed. L. Nadel and W. Sinnott-Armstrong, 173-88. New York: Oxford University Press.

Nahmias, E. 2002. When consciousness matters: A critical review of Daniel Wegner's The illusion of conscious will. *Philosophical Psychology* 15(4): 527-541.

Nahmias, E. 2011. 2011. Intuitions about free will, determinism, and bypassing. In R. Kane (Ed.), *The Oxford handbook of free will, 2nd ed.*, pp. 555-576. New York: Oxford University Press.

Nichols, S. 2004. The folk psychology of free will: Fits and Starts. *Mind and Language* 19: 473-502.

Nisbett, R., and T. Wilson. 1997. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231-58.

O'Connor, T. 1995. Agent causation. In *Agents, causes and events: Essays on free will and indeterminism*, ed. Timothy O'Connor, 173-200. New York: Oxford University Press.

O'Connor, T. 2000. *Persons and causes: The metaphysics of free will*

Obhi, S.S. and P. Haggard. 2004. Free will and free won't: Motor activity in the brain precedes our awareness of the intention to move, so how is it that we perceive control? *American Scientist* 92 (July-August): 358-65.

Oshana, M. 1997. Ascriptions of responsibility. *American Philosophical Quarterly* 34: 71-83.

Pereboom, D. 2001. *Living without free will*. Oxford: Cambridge University Press.

Pereboom, D. 2014. *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Pereboom, D. and G. D. Caruso. 2017. Hard-Incompatibilism Existentialism: Neuroscience, Punishment, and Meaning in Life. In *Neuroexistentialism: Meaning, morals, and purpose in the age of neuroscience,* eds. Gregg D. Caruso and Owen Flanagan. New York: Oxford University Press.

Pink, T. 2009. Free will and consciousness. In *The Oxford companion to consciousness*, ed. Timothy Bayne, Alex Cleeremans, and Patrick Wilken, 296-300. New York: Oxford University Press.

Pockett, S. 2004. Does consciousness cause behavior? *Journal of Consciousness Studies* 11: 23-40.

Rosenthal, D. 2002. The Timing of Conscious States. *Consciousness and Cognition* 11 (2): 215-220.

Rosenthal, D. 2005. *Consciousness and mind*. New York: Oxford University Press.

Ross, P. 2006. Empirical constraints on the problem of free will. In *Does consciousness cause behavior?* Eds. Susan Pockett, William P. Banks, and Shaun Gallagher, 125-44. Cambridge, MA: MIT Press.

Scanlon, T. 1998. *What we owe to each other*. Cambridge: Harvard University Press.

Schopp, R.F. 1991. *Automatism, insanity, and the psychology of criminal responsibility: A philosophical inquiry.* Cambridge: Cambridge University Press.

Searle, J. 2000. Consciousness, free action and the brain. *Journal of Consciousness Studies* 7 (10): 3-22.

Searle, J. 2001a. *Rationality in action*. Cambridge, MA: MIT Press.

Searle, J. 2001b. Free will as a problem in neurobiology. *Philosophy* 76: 491-514.

Searle, J. 2007. *Freedom and neurobiology: Reflections on free will, language and political power.* Columbia University Press.

Shepherd, J. 2012. Free will and consciousness: Experimental studies. *Consciousness and Cognition* 21: 915-927.

Shepherd, J. 2015a. Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology* 28 (7): 929-946.

Shepherd, J. 2015b. Scientific challenges to free will and moral responsibility. *Philosophy Compass* 10(3): 197-207.

Sher, G. 2009. *Who knew? Responsibility without awareness*. New York: Oxford University Press.

Shoemaker, D. 2011. Attributability, answerability, and accountability: Toward a wide-theory of moral responsibility. *Ethics* 121 (3): 602-632.

Siddiqui, F., E. Osuna, and S. Chokroverty. 2009.Writing emails as part of sleepwalking after increase in zolpidem. *Sleep Medicine* 10: 262-64.

Sie, Maureen, and Arno Wouters. 2010. The BCN challenge to compatibilist free will and personal responsibility. *Neuroethics* 3 (2): 121-33.

Smith, A. 2005. Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115: 236-271.

Smith, A 2008. Control, responsibility, and moral assessment. *Philosophical Studies* 138: 367-92.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11 (5): 543-45.

Strawson, G. 1986. *Freedom and belief.* Oxford: Oxford University Press. Revised edition 2010.

Strawson, G. 1994. The impossibility of moral responsibility. *Philosophical Studies* 75 (1): 5-24.

Taylor, R. 1992. *Metaphysics*. 4th Edition. Englewood Cliffs, NJ: Prentice-Hall.

Tegmark, M. 1999. The importance of quantum decoherence in brain processes. *Physics Review E* 61: 4194-206.

Uhlmann, E.L., and G.L. Cohen. 2005. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16: 474-480.

Waller, B. 2011. *Against moral responsibility*. Cambridge, MA: MIT Press.

Watson, G. 1996. Two faces of responsibility. *Philosophical Topics* 24 (2): 227-248.

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

Weingarten, E., Q. Chen, M. McAdams, J. Yi, J. Hepler, and D. Albarracin. 2016. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin* 142 (5): 472-497.

Wilson, T. 2002. *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: The Belknap Press of Harvard University Press.

Wood, L. 1941. Determinism: Free will is an illusion. *Philosophy* 16: 386-89.

Wolf, S. 1990. *Freedom and Reason*. Oxford: Oxford University Press.