Consciousness, Free Will, and Moral Responsibility

Gregg D. Caruso

Forthcoming in *Routledge Handbook of Consciousness*.
Edited by Rocco J. Gennaro [Estimated 2018]

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? For example, 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? Clarifying the relationship between consciousness and free will is imperative if we want to evaluate the various arguments for and against free will.

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 three 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.

(2) The second category also maintains that consciousness is a necessary condition for free will, but believes 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, 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. Classical compatibilism, for example, typically focused on the correct semantic analysis of the expression "could have done otherwise," without any reference to consciousness or experience. More recently, 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 reply 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.

I. Free Will and Moral Responsibility

Before discussing each of the categories in detail, let me begin by defining what I mean by free will and moral responsibility. The concept of *free will*, as it is typically understood in the contemporary debate, is a term of art referring to the control in action required for a core sense of moral responsibility. This sense of moral responsibility is traditionally 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; Caruso and Morris 2016). 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.

Historically, the problem of free will has centered on *determinism*—the thesis that every event or action, including human action, is the inevitable result of preceding events and actions and the laws of nature. *Hard determinists* and *libertarians* argue that causal determinism is incompatible with free will—either because it precluded the *ability to do otherwise* (leeway incompatibilism) or because it is inconsistent with one's being the "ultimate source" of action (source incompatibilism). The two views differ, however, on whether or not they accept determinism. Hard determinists claim that determinism is true and hence no free will, while

libertarians reject determinism and defend an indeterminist conception of free will. *Compatibilists*, on the hand, attempt to reconcile determinism and free will. 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.

More recently a new crop of *free will skeptics*—i.e., those who doubt or deny the existence of free will—has emerged that are agnostic about the truth of determinism. Most argue 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). What all these arguments for free will skepticism have in common is the claim that what we do, and the way we are, is ultimately the result of factors beyond our control and because of this we are never morally responsible for our actions in the *basic desert* sense.

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 (Libet et al. 1993; Soon et al. 2008), and recent findings in psychology and social psychology on automaticity, situationism, and the adaptive unconscious (Nisbet and Wislon

1997; Bargh 1997; Bargh and Chartrand 1999; Bargh and Ferguson 2000; Doris 2002, Wilson 2002). 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. 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.

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 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 can be reconciled with determinism, chance, or luck, 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., Sie and Wouters 2010, Nadelhoffer 2011; King and Carruthers 2012; Caruso 2012, 2015b; Levy 2014).

II. Is Consciousness Necessary for Free Will?

Turning now to the relationship between consciousness and free will, the three 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), Gregg Caruso (2012, 2015b), and Joshua Shepherd (2012, 2015), 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 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). 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) influential 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 Shepherd (2015) and Levy (2014) have noted, such accounts typically impart no significant role to consciousness. Indeed, Gideon Yaffe claims that "there is no reason to suppose that consciousness is required for reasons-responsiveness" (2012: 182). Given this, reasons-responsive accounts can also 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 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, in his view,

inadequate understanding of it. 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. The basic idea behind Sher's positive 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, however, is different than Sher's since she attacks the notion that voluntariness (or active control) is a precondition of moral responsibility rather than the epistemic condition. She writes, "our commonsense intuitions do not, in fact, favor a volitionalist criterion of responsibility, but a rationalist one." That is to say, "the kind of activity implied by our moral practices is not the activity of [conscious] choice, but the activity of evaluative judgment." She argues that 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 (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.

Contrary to these views, Neil Levy (2014), Joshua Shepherd (2012, 2015), and Gregg 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.

Levy further 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. 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 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, 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).

Second, on Levy's account, information of the right kind must be personally available to ground moral responsibility. But what kind of information is the right kind? 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. 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.

I should note that in *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. 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. 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. In a series of studies he provides compelling evidence that ordinary folk accord a central place to consciousness when it comes to free will and moral responsibility—furthermore, "the way in

which it is central is not captured by extant [Real or] Deep Self Views" (2015: 938). For details, see Shepherd (2012, 2015).

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. 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.

Libet's findings, in conjunction with additional findings by John Dylan Haynes (Soon et al. 2008) and Daniel Wegner (2002), 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. 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 can still be causes.

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) 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% above 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).

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.

But even if neuroscience is not able to refute free will, there are other empirical threats to free will and moral responsibility that remain. And these threats challenge a different sort of consciousness thesis—the one proposed by Neil Levy. In fact, Levy argues that 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. 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.

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*. 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 sleepwalking 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. Siddiqui et al. (2009), for example, recently described a case of sleep emailing. These cases illustrate the complexity of the behaviors in which agents may engage in the apparent absence of awareness. Levy argues that such behaviors 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 accounts maintain.

Assuming that Kenneth 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, "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)" (Levy 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.

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 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; for a reply, see Caruso 2015b).

III. Volitional Consciousness

Let me end by noting one last category of views—i.e., those that maintain that consciousness is necessary condition for free will and that the condition can be satisfied. Due to space, I will limit my discussion to two leading libertarian accounts of volitional consciousness, those of John Searle and David Hodgson

Both Searle (2000, 2001) and Hodgson (2005, 2012) 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 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 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, 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 makes clear—*Rationality + Consciousness = Free Will* (2012). 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.

There are, however, several important challenges confronting libertarian accounts of volitional consciousness. 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; Klein et al. 2002). Second, work by Daniel Kahneman (2011), Jonathan Haidt (2001, 2012), and others (e.g., Wilson 2002) has shown that much of what we take to be "unbiased conscious deliberation" is at best rationalization. Third, 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. This appeal to quantum mechanics to account for conscious rational behavior, however, 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. 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. 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 it is motivated 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). Well, positing one mystery to account for another will likely be unconvincing to many.

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 2015c). The core objection is that because 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 that go beyond the naturalistic commitments of Searle and Hodgson, 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 three 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 will is imperative if one wants to evaluate the various arguments for and against free will.

References

Arplay, N. (2002) *Unprincipled Virtues: An Inquiry into Moral Agency*, New York: Oxford University Press.

Arplay, N., and T. Schroeder. (1999) "Praise, 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 R. S. Wyer, Jr. (ed.) *The Automaticity of Everyday Life: Advances in Social Cognition,* Vol. 10, Mahwah, NJ: Erlbaum.

Bargh, J.A., and T.L. Chartrand (1999) "The Unbearable Automaticity of Being," *American Psychologist* 54 (7): 462-79.

Bargh, J.A., and M.J. Ferguson (2000) "Beyond Behaviorism: On the Automaticity of Higher Mental Processes," *Psychological Bulletin* 126 (6): 925-45.

Baumeister, R.F. (2008) "Free Will in Scientific Psychology," *Perspectives of Psychological Science* 3 (1): 14-19.

Caruso, G.D. (2005) "Sensory States, Consciousness, and the Cartesian Assumption," in N. Smith and J. Taylor (eds.) *Descartes and Cartesianism*, UK: Cambridge Scholars Press.

Caruso, G.D. (2012) *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*, 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): 21-26.

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.

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.

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., and M. Ravizza (1998) *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

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. (2005) "A Plain Person's Free Will," *Journal of Consciousness Studies* 12 (1): 1-19.

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.

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: 200-28.

Klein, S., K. Rozendal, and L. Cosmides (2002) "A Social-Cognitive Neuroscience Analysis of the Self," *Social Cognition* 20 (2): 105-35.

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.

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 R. Kane (ed.) *The Oxford Handbook of Free Will*, 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: 274-293.

Mele, A. (2009) *Effective Intentions*, New York: Oxford University Press.

Nadelhoffer, T. (2011) "The Threat of Shrinking Agency and Free Will Disillusionism," in L. Nadel and W. Sinnott-Armstrong (eds.) *Conscious Will and Responsibility: A Tribute to Benjamin Libet*, New York: Oxford University Press.

Nahmias, E. (2011) "Intuitions about Free Will, Determinism, and Bypassing," in R. Kane (ed.), *The Oxford Handbook of Free Will, 2nd ed.*, New York: Oxford University Press.

Nisbett, R., and T. Wilson (1997) "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84: 231-58.

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 G.D. Caruso and O. Flanagan (eds.) *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience,* New York: Oxford University Press.

Rosenthal, D. (2005) *Consciousness and Mind*, New York: Oxford 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. (2001) *Rationality in Action*, Cambridge, MA: MIT Press.

Searle, J. (2007) *Freedom and Neurobiology: Reflections on Free Will, Language and Political Power*, New York: Columbia University Press.

Shepherd, J. (2012) “Free Will and Consciousness: Experimental Studies,” *Consciousness and Cognition* 21: 915-927.

Shepherd, J. (2015) “Consciousness, Free Will, and Moral Responsibility: Taking the Folk Seriously,” *Philosophical Psychology* 28 (7): 929-946.

Sher, G. (2009) *Who Knew? Responsibility Without Awareness*, New York: Oxford University Press.

Siddiqui, F., E. Osuna, and S. Chokroverty (2009) “Writing Emails as Part of Sleepwalking After Increase in Zolpidem,” *Sleep Medicine* 10: 262-64.

Sie, M., and A. 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, C.S., M. Brass, H-J. Heinze, and J-D. 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.

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.

Wegner, D. (2002) *The Illusion of Conscious Will*, Cambridge, MA: MIT Press.

Wilson, T. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious*, Cambridge, MA: The Belknap Press of Harvard University Press.

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

Yaffe, G. (2012) “The Voluntary Act Requirement,” in M. Andrei (ed.) *Routledge Companion to Philosophy of Law*, New York: Routledge.