When Do Robots Have Free Will? Exploring the Relationships between (Attributions of) Consciousness and Free Will

Eddy Nahmias, Corey Allen, and Bradley Loveall
Georgia State University
enahmias@gsu.edu

****prepublication draft of chapter in *Free Will, Causality, and Neuroscience* (edited by Bernard Feltz, Marcus Missal and Andrew Cameron Sims, Brill Publishers).**

1. The often implicit, yet essential, connection between consciousness and free will

Imagine that, in the future, humans develop the technology to construct humanoid robots with very sophisticated computers instead of brains and with bodies made out of metal, plastic, and synthetic materials. The robots look, talk, and act just like humans and are able to integrate into human society and to interact with humans across any situation. They work in our offices and our restaurants, teach in our schools, and discuss the important matters of the day in our bars and coffeehouses. How do you suppose you'd respond if you were to discover one of these robots attempting to steal your wallet or insulting your friend? Would you regard them as free and morally responsible agents, genuinely deserving of blame and punishment?

If you're like most people, you are more likely to regard these robots as having free will and being morally responsible if you believe that they are conscious rather than non-conscious. That is, if you think that the robots *actually experience* sensations and emotions, you are more likely to regard them as having free will and being morally responsible than if you think they simply behave like humans based on their internal programming but with no conscious experiences at all. But *why* do many people have this intuition? Philosophers and scientists typically assume that there is a deep connection between consciousness and free will, but few have developed theories to explain this connection. To the extent that they have, it's typically via some cognitive capacity thought to be important for free will, such as reasoning or deliberation, that consciousness is supposed to enable or bolster, at least in humans. But this sort of connection between consciousness and free will is relatively weak. First, it's contingent; given our particular cognitive architecture, it holds, but if robots or aliens could carry out the relevant cognitive capacities *without* being conscious, this would suggest that consciousness is not constitutive of, or essential for, free will. Second, this connection is derivative, since the main connection goes through some capacity other than consciousness. Finally, this connection does not seem to be focused on *phenomenal* consciousness (first-person experience or qualia), but instead, on *access* consciousness or self-awareness (more on these distinctions below).

Perhaps the most substantive claims about the necessity of consciousness for free will come from scientists who discuss free will. For instance, 'willusionists', who say that scientific

research suggests that free will is an illusion, typically reach that conclusion by assuming that free will requires that one's 'conscious will' causes one's actions. Since willusionists argue that scientific research shows that conscious will does *not* cause our actions, they conclude that free will is an illusion (see, e.g., Libet 1999, Wegner 2002, Bargh 2008, and Harris 2012). Willusionists support their conclusion by arguing that research in neuroscience and psychology suggests that conscious mental states and processes do not play a causal role in decisions and actions, because non-conscious neural or psychological processes happen first.[1] These scientists, however, do not say much about *why* consciousness is crucial for free will. They typically assert the essential connection with claims such as Roy Baumeister's: "If there are any genuine phenomena associated with the concept of free will, they most likely involve conscious choice" (2008, 76; see also Libet 1999 and Wegner 2002).

Philosophers tend to agree that consciousness is necessary for free will. For instance, when they respond to willusionists, they typically dispute the relevance of the neuroscientific studies, the dualist or libertarian stipulations about how to define free will, or *which* conscious mental states are relevant to free choices (e.g., important deliberations, not consciousness of an intention to move a moment before an inconsequential movement like pressing a button). But these philosophers do not reject the importance of consciousness for free will. For example, Al Mele notes that "[i]f all behavior were produced *only* by nonconscious processes, and if conscious decisions (or choices) and intentions (along with their physical correlates) were to play no role at all in producing any corresponding actions, free will would be in dire straits" (2010, 43). And Eddy Nahmias suggests that: "Free will requires that one's actions properly derive from reasons for action that one has at some point consciously considered (or at least that one would accept if one considered them)" (2014, 18).

But it did not take scientists challenging the role of consciousness for philosophers to suggest that it is required for free will and moral responsibility. For instance, Galen Strawson writes, "To be responsible... one must have consciously and explicitly chosen to be the way one is, mentally speaking, in certain respects" (1994, 6). Randy Clarke writes, "Free will requires a capacity for rational self-determination… a free agent must be able to exercise [this capacity] consciously … an agent who is not even capable of conscious, effective practical reasoning does not have the variety of rational control in acting that we prize" (2003, 16). And Isaiah Berlin writes, "I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes that affect me, as it were, from outside" (1958, 203). Across diverging theorists—from compatibilists to libertarians to skeptics about free will—one truth seems to be self-evident: that free will requires consciousness.

Yet, despite the fact that this free will-consciousness connection is so pervasive among scientists and philosophers, the connection has typically been asserted without much explanation or defense, often taken for granted or left implicit. As Gregg Caruso points out, this is an

[1] See Mele (2010) and Nahmias (2014) for various responses to the evidence willusionists use, their interpretation of its relevance to conscious intention formation and to free will, and to the definition of free will the willusionists (mistakenly) assert as both dominant in philosophy and as commonsensical.

explanatory gap that must be filled: "Clarifying the relationship between consciousness and free will is imperative if we want to evaluate the various arguments for and against free will" (2016, 78). It may be that the connection is under-analyzed precisely because it is so intuitive that we tend not to notice it. We've never encountered agents that seem autonomous and responsible which we do not also assume to be conscious. But perhaps that will change if we develop autonomous robots (or meet intelligent aliens) and we are unsure about their consciousness. Furthermore, perhaps we can learn more about the free will-consciousness connection by exploring ordinary people's reactions to such possibilities and trying to tease apart which features of consciousness underlie their attributions of free will.

Indeed, until recently, little attention has been paid to ordinary people's attitudes about the connection between consciousness, free will, and moral responsibility. Given that many free will theorists appeal to commonly held intuitions as evidence for their theory, it is important that philosophical theorizing about concepts such as free will track ordinary understanding of those conceptions, or conversely, provide an error theory to explain why those intuitions are mistaken (see, e.g., Murray and Nahmias 2014). While some experimental philosophers and psychologists have conducted studies on people's intuitions and attitudes about free will and moral responsibility,[2] the relationship between attributions of free will and consciousness has been largely underexplored.

Recognizing this gap in the literature, Joshua Shepherd conducted a series of studies designed to understand people's attitudes about the role that consciousness plays in grounding free will and moral responsibility (e.g., 2012, 2015). Across several studies, Shepherd finds that people are much more likely to judge an agent to be free and responsible if the agent is conscious and to judge that an agent's particular actions are free and responsible when they are carried out consciously rather than non-consciously.

In one intriguing study, Shepherd asked participants to imagine the existence of sophisticated humanoid robots who "look, talk, and act just like humans, and they integrate into human society with no problem at all" (2015, 939). Some participants read scenarios that describe these robots as possessing consciousness: "They actually *feel* pain, *experience* emotions, *see* colors, and *consciously* deliberate about what to do"; while other participants read that the robots are *not* conscious: "they do not *actually feel* pain … they do not *experience* emotions, they do not *see* colors, and they do not *consciously* deliberate about what to do" (939). Some participants read a scenario in which one of these robots, Sal, steals a wallet he finds, while other participants read a scenario in which Sal returns a wallet he finds. Across scenarios describing both the bad and good action, participants who were told that the robots were conscious tended to judge Sal to be free, blameworthy (or praiseworthy), and morally responsible, while those who were told that the robots were not conscious tended *not* to attribute free will or responsibility to Sal (940). Shepherd concludes that these results show that most people believe that conscious states and processes play a central role in grounding free will and

[2] See, e.g., Nahmias, Morris, Nadelhoffer, & Turner, 2006; Nichols & Knobe, 2007; Monroe, Dillon, & Malle, 2014; Stillman, Baumeister, & Mele, 2011; Vonasch, Baumeister, & Mele, 2018.

moral responsibility. Shepherd speculates about some reasons people may make the connection between free will and consciousness, and how some philosophical theories might align with or revise ordinary intuitions about the connection. He concludes that his findings suggest that philosophers should "either develop a substantive theory of the connection between consciousness on the one hand and free will and moral responsibility on the other, or offer justification for jettisoning this seemingly central part of our commonsense understanding of free will and moral responsibility" (944).[3]

Here, we describe studies we conducted that build off of Shepherd's studies, and we take up his explanatory challenge. In fact, we hope to use our studies to begin to distinguish *which* features of consciousness, or the capacities they might allow, people see as most essential for free will, and *why* these features or capacities are especially relevant or essential. Our aim is to bring to the surface implicit connections that might underlie the strong intuition among most people—including most philosophers and scientists who discuss free will—that the capacity to have conscious experiences is crucial for free will and responsible agency. If so, it might be that philosophers can even develop plausible theories by drawing on the connections underlying ordinary thinking. In any case, we'll try to develop one such theory that we take to be plausible.

2. Some potential connections between consciousness and free will

There are several plausible features of consciousness that might be considered relevant to free will and moral responsibility. One historically prominent route emphasizes the phenomenology of free will. For example, Jean-Paul Sartre (1943) suggested that being conscious (specifically self-conscious) is necessarily involved in being radically free. Others have argued that the experience of free will is necessary for having free will. Galen Strawson, for instance, writes, "But why should lack of explicit awareness of [freedom] be supposed to have as a consequence lack of [freedom] itself? ... Well, that is the question. But it does seem to be so" (1986, 307). The idea may be that the first-person experience of having open alternatives for future choices is essential for actually possessing free will, but it's not clear *why* the experience of freedom is necessary. For example, must the phenomenal experience of freedom play some *causal* role in one's decisions and actions, or could the experience be epiphenomenal? If a causal role is required, then the question is what the experience of freedom is causing and what agents lack if they can behave in similar ways without the experience of freedom. If the experience plays no causal role, then it is even more mysterious what role it plays in making the agent free or morally responsible.

Another feature of consciousness that might be relevant to free will is its role in grounding libertarian free will. One might defend this view in a few ways. Perhaps, for example, consciousness bears some relation to a non-physical mind or soul that can make free choices and causally interact with the physical brain and body (e.g., Swinburne 2013). The idea seems to be

[3] For studies on attributions of consciousness, see, e.g., Huebner, 2010; Jack & Robbins, 2012; Knobe & Prinz, 2008.

that the conscious self can be an uncaused cause, free from the deterministic chain of cause and effect in the physical world. In addition to the mysterious causal interaction between non-physical minds and physical bodies that this view suggests, it also does not explain *why* it is the mind or soul's capacity for *consciousness* that allows it to be an uncaused cause. Other libertarians have connected consciousness to free will via quantum theory, gesturing towards the indeterminism of the dominant theory of quantum mechanics or towards the alleged role that consciousness plays in collapsing the wave function (see, e.g., Penrose 1991, Stapp 2001, and Hodgson 2002). At this stage, however, these views seem to try to solve the mystery of (libertarian) free will by conjoining the mystery of quantum physics with the mystery of consciousness. Robert Kane, offering a naturalistic libertarian view, suggests that consciousness may allow the unity of the self: "it may be that both the unity of conscious experience and the unity of the self-network are somehow related to the quantum character of reality" (1998, 195). It is plausible that free will requires a unified self and that we have conscious experiences of being a unified self at a time and across time (an experience some Humeans and Buddhists would say is illusory), and below we suggest there are specific features of consciously caring that are relevant to demarcating the self. It is unclear, however, why the unity of self should be associated specifically with a *libertarian* theory of free will.

Indeed, some compatibilists about free will and determinism suggest that consciousness is relevant because it allows the integration of information such that the agent has the ability to access her competing reasons and values during deliberation and make choices that represent her (unified) self. For example, Neil Levy argues for the 'consciousness thesis' which says that "consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility (2014, 1).[4] However, like some other compatibilist theories described below, Levy's does not focus on *phenomenal* consciousness (qualia or the sensory and emotional qualities associated with conscious experiences). Rather, Levy focuses specifically on *access* consciousness. Information is 'access conscious' to an agent when it is accessible for use by a wide range of cognitive systems, including deliberation, reasoning, and verbal report (Block 1995). It's controversial whether the distinction between these concepts of consciousness maps smoothly onto human psychology or fits with ordinary people's understanding of consciousness, but Levy uses it to focus on the importance of access consciousness, suggesting that phenomenal consciousness is irrelevant.

Another prominent compatibilist theory may similarly suggest that, for the freedom associated with moral responsibility, it is accessibility of information to reasoning processes that matters more than phenomenal or qualitative experience. Reasons-responsive theories emphasize the control that access consciousness enables over one's decisions and actions as a result of reasoning, deliberation, and self-reflection. On Fischer and Ravizza's (1998) account, for example, agents are morally responsible for actions that are caused by moderately reasons-

[4] Note that Levy focuses here on moral responsibility, not free will, and that he is a somewhat non-standard compatibilist, in that he thinks determinism does not rule out free will or moral responsibility, but he thinks we have neither because of an argument from luck.

responsive mechanisms. As Shepherd (2015, 943) points out, however, reasons-responsive theories often emphasize features of decision-making that are unrelated to conscious experience. Indeed, on some interpretations of reasons-responsive compatibilism, it's unclear whether agents need to have conscious experiences in order to have the capacity for free will and moral responsibility (e.g., Yaffe 2012, 182). While access-conscious mental states are certainly required on Fischer and Ravizza's view, it's not entirely clear what role, if any, *phenomenal* mental states and processes play in enabling free will and moral responsibility. Though Fischer and Ravizza argue that free agents must 'take ownership' of the relevant mechanisms, they don't say whether or why these agents must be conscious of the practical reasoning processes that they carry out.

Another type of compatibilist account suggests that free will involves decisions and actions caused by the 'deep self' or 'real self' as labeled by Susan Wolf (1990). She is referring to theories that pick out freely willed actions as the ones caused by those (first-order) desires or motivations that the agent (second-order) desires to move her (Frankfurt 1971), or that she identifies herself with (Frankfurt 1987), or that accord with her considered values (Watson 1975). These theories seem to require that the agent has free will only if she acts on motivational states that she is consciously aware of and consciously endorses. If so, they also seem to link free will to access consciousness or a type of self-reflective awareness. One might wonder whether such higher-order representational states require phenomenal consciousness—whether there must be anything it is like to experience them—or whether a sophisticated robot (or even humans in some instances) could carry out such higher-order representation without any phenomenology at all. As Daniel Dennett suggests in describing such robots: "our imagined [non-conscious] creatures would be equally able to engage in rational self-evaluation. They would be equipped to react appropriately when we represent reasons to them. Isn't that what freedom hinges upon, whether or not it amounts to consciousness?" (1984, 43).

Yet another type of compatibilist theory, related to these deep self views, suggests that consciousness is *not* required for free and responsible actions. These 'quality of will' (or self-expression) theories say that agents are responsible for those actions that express the agent's concern or consideration of others (their quality of will), which is sometimes identified as actions expressing the agent's deep self (see, e.g., Arpaly 2003, Smith 2005, Sher 2009, Buss 2012, and Sripada 2016). These theorists have argued that actions can express an agent's quality of will even when motivated by values that the agent consciously rejects (e.g., Huck Finn protecting Jim even though he thinks he should not) or in cases of negligence, when the agent, for instance, does not care enough to show up to help his friend move but did not *consciously* try to forget. These instances of responsibility for specific non-conscious actions are plausible. However, it's unclear whether these theorists would go on to argue that it's irrelevant to free will and responsibility whether a creature is phenomenally conscious at all (sometimes called 'creature consciousness').

Indeed, most quality of will theorists take their inspiration from Peter Strawson, who lays the foundation for the proposal that we will suggest for the connection between phenomenal consciousness and free will. Strawson (1962) argues that freedom and responsibility are

grounded in our reactive attitudes, such as indignation, gratitude, and guilt, that we express in interpersonal relationships. According to Strawson, agents are morally responsible when they are apt targets of these reactive attitudes—for instance, when it is appropriate to feel resentment towards them when their actions express a poor quality of will towards you, to feel gratitude towards them when their actions express a good quality of will towards you, and to feel guilt when your own action expresses poor quality of will towards others. As such, Strawson's account ties free and responsible agency to the capacity to experience and express certain moral emotions, and it suggests that we attribute such agency only to other agents whom we perceive as feeling relevant emotions and expressing them in their actions. Hence, on a Strawsonian account, it might be the ability to consciously experience emotions that bridges phenomenal consciousness and the freedom required for responsibility.

A related view suggests that free and responsible agency is tied to our ability to *care* about what motivates us. On this view, actions expressing our deep self or quality of will are those that are caused by what we care about. For instance, Harry Frankfurt modifies his earlier views, which focused on higher-order desire and identification, to focus on the role that caring plays for grounding agency. He writes that a free agent is "prepared to endorse or repudiate the motives from which he acts … to guide his conduct in accordance with what he really cares about"; and he adds that what is essential to freedom pertains to "what a person cares about, what he considers important to him… Caring is essentially volitional… similar to the force of love" (1999, 113-114). Chandra Sripada develops these ideas, arguing that one is morally responsible for an action only when it expresses one's deep self, and that the actions that express one's self are precisely those motivated by one's cares (2016). He defines cares in terms of what functional role they play in our psychology: they are foundational motives—i.e., intrinsic and ultimate, such that many of our other desires motivate actions that aim at satisfying our cares—and we desire to maintain our cares, and feel a sense of loss when we alter them.

Sripada's conative account is contrasted with the more cognitive deep self approaches described above that seem to require access (or self-) consciousness. His account suggests a more important role for *phenomenal* consciousness, because it is directly tied to emotion. He writes, "caring is also associated with a rich and distinctive profile of emotional responses that are finely tuned to the fortunes of the thing that is the object of the care" (2016, 1210).[5] For instance, if Paul *cares* about the plight of Sudanese children, then "if the fortunes of the Sudanese children are set back, Paul is susceptible to sadness, disapprobation, and despair" (1230-31). Now, Sripada and other theorists writing about cares do not explicitly point out that phenomenal consciousness is crucial for an agent to be able to have cares, likely because they take it to be understood that feeling emotions like sadness, despair, and joy requires the ability to have phenomenally conscious experiences. Indeed, it is difficult for us to imagine creatures (such as humanoid robots) that lack conscious experiences entirely yet are also able to have the sort of

[5] Sripada cites his debt to David Shoemaker's excellent paper on caring and agency (2003). Shoemaker writes, "the relation between cares and affective states is extremely tight" (93) and "the emotions we have make us the agents we are" (94).

emotional responses required for them to *care* about what they do or what happens to them. They might have motivational states, they might represent them and evaluate them, but they would not seem to have the capacities to feel the sort of satisfaction or suffering that are constitutive of caring.

We have now seen two related accounts that situate certain emotions at the heart of free and responsible agency, Strawsonian accounts based on reactive attitudes and self-expressive accounts that focus on the capacity to care.[6] We propose that these connections provide the link between free will and (specifically) phenomenal consciousness. Then, we offer some initial evidence that people's intuitive understanding of free will points towards this proposal.

3. Emotional experiences as the essential link between consciousness and free will

As we've seen, different theories suggest different connections between free will and consciousness, and the connection might be more or less direct and it might be considered contingent or conceptually necessary. Some accounts (e.g., some libertarian theories) suggest that the connection is direct and conceptual, such that free will, *by definition*, requires consciousness of some sort. More often, the connection, to the extent it is discussed at all, takes a less direct route and suggests a contingent relationship. The idea is that free will requires something *x*, like control, self-awareness, or reasoning, and that *x* is what requires consciousness of some sort, at least given humans' particular cognitive architecture. For example, a reasons-responsive compatibilist might argue that free will requires certain deliberative capacities which happen to require, in some cases, conscious processes in creatures like us. However, on such views, it is unclear whether consciousness, especially phenomenal rather than access consciousness, is necessary for free will or whether it is only contingently related to free will in virtue of the fact that it enables these deliberative capacities that are themselves required for free will. Perhaps, for example, some other complex cognitive agent could carry out the behaviors that are enabled by deliberative capacities *without* phenomenal consciousness.

Again, this suggests that we might be able to test the free will-consciousness connection by considering robots or aliens that are stipulated to have the relevant capacity *x* and behave just like humans but to do so without phenomenal consciousness. To the extent that such creatures are conceivable, we might wonder whether they have free will. If such creatures still plausibly *have* free will, then it suggests a more indirect, contingent relationship between consciousness and free will. However, if the creatures plausibly *lack* free will, even though they behave like humans, it suggests that there must be some more direct relationship, such that the capacity of a creature to be phenomenally conscious is constitutive of or essential for that creature to have free will.

[6] Shepherd and Levy (forthcoming) briefly suggest another idea in this ballpark. They posit that the moral knowledge required to be a responsible agent requires moral perception which requires phenomenal consciousness in order to understand the intrinsic moral value of one's own and others' experiences of pleasure and pain.

We suggest a relatively direct connection between phenomenal consciousness and free will. The basic idea is that one thing that matters when it comes to being a free agent is that things can really *matter* to the agent. Moreover, in order for anything to matter to an agent, she has to be able to experience the negative and positive consequences of her choices, to be able to feel pain, suffering, and disappointment for choices whose outcomes conflict with what she cares about, and to feel pleasure, joy, and satisfaction for choices whose outcomes sustain her cares, and plausibly to foresee experiencing these feelings when evaluating options for future action. Feeling pain and pleasure, and emotions such as anxiety and joy, requires phenomenal consciousness. These mental states are essential for caring about anything. As Sripada suggests, when someone cares about something, "Her emotions are bound tightly to the fortunes of the thing… These observations suggest that there is a basic *conceptual* tie between the syndrome of dispositional effects [the functional roles] associated with cares and *what it is* for something to matter to a person" (2016, 1211). Furthermore, when it comes to consequential or moral decisions involving interpersonal relations, it is essential that the agent can also experience the Strawsonian emotions such as shame, pride, regret, gratitude, and guilt. After all, many of our deepest cares involve other people, how our actions affect them and how their actions affect us. So, on this view, the connection between free will and consciousness goes through the capacities to feel emotions that ground mattering, caring, and reactive attitudes.

This view suggests that it is implausible for anything to *really* matter to an agent that cannot consciously feel anything, even if that agent were sophisticated and intelligent enough to behave just like us. However, such a robot or alien may be able to *behave* much like us and have the capacities for intelligent action, the evaluation of options, and even complex reasoning, without having phenomenal consciousness. If so, it seems nothing would really matter to such a creature, that it would not really care about what decisions it made. And it seems—to us at least—that it would lack free will.

This, then, is the intuitive connection between consciousness and free will that we wanted to test and compare to other potential connections, motivated by the pervasive implicit or explicit claims about the consciousness-free will connection, by the relative paucity of explanations for it, and by Shepherd's initial work on this topic.

4. Studies on attributions of consciousness and free will

Following up on Shepherd's paradigm, we designed two studies to explore people's attributions of consciousness and free will to humanoid robots. The goal was to try to have people consider creatures that look and act like humans while avoiding people's default and implicit attributions of free will and consciousness to humans (and perhaps to any similar biological creatures). As we will see, most people seem to have implicit representations of robots as non-conscious and unfree. (In current studies not described here, we use scenarios that describe alien lifeforms, and most people seem to have implicit representations of these aliens, as well as many animals, as both conscious and having free will.)

4.1. Study 1: Participants and Design

Our first study had 373 participants (68.2% female, 31.8% male; mean age 19.78, ranging from 18-38), who were undergraduates at Georgia State University. After removing 49 for incomplete data, for missing attention checks, or for spending too little time on the survey, the sample size was 324. This study (as well as its follow-up) was approved by the university's Institutional Review Board, and participation was voluntary and conditioned on informed consent. The study was administered online via Qualtrics.

This experimental vignette study used a between-group design with random assignment to two Learning conditions, as well as to a third control condition. The Learning manipulation varied whether the humanoid robots were able to behave like humans because they could learn based on experiences or because they were preprogrammed with all necessary knowledge. The control condition did not include robots, but instead discussed humans. All scenarios end with a paragraph describing a variety of behaviors that would typically be interpreted as involving conscious experiences in humans, such as feeling cold, expressing empathy, or making a decision. The primary dependent measures consisted of responses to individual statements (from 1 - "Strongly Disagree" to 7 - "Strongly Agree") that were summed together to create sub-scale composite items representing the attributions of the following capacities to either the robots or humans: free will, moral responsibility, basic emotions, Strawsonian emotions, conscious sensation, reason and reflection, and theory of mind.

The experimental vignettes read as follows:

In the future, humans develop the technology to construct humanoid robots with very sophisticated computers instead of brains and with bodies made out of metal, plastic, and synthetic materials. The robots look, talk, and act just like humans and are able to integrate into human society and to interact with humans across any situation. The only way to tell if they are a humanoid robot instead of a human being is to look inside of them (using x-ray, for example).

(Learning Condition) These robots have various components that process information and allow them to learn from their interactions so that they can change over time. For example, like humans they are able to learn new languages by interacting with people using those languages. Their ability to learn allows the robots to adapt to different situations.

(Pre Programmed Condition) These robots have various components that were pre-programmed with all of the information they would need to behave appropriately in any situation. For example, unlike humans, they are pre-programmed to be able to speak any language when interacting with people using the language. Their program allows the robots to behave appropriately across different situations.

Imagine you are asked to observe some of these robots over the course of several weeks, and you see different robots carrying out a wide range of behaviors. For instance, one of the robots, Taylor, gets a hand slammed in a car door and Taylor yells out, grabs the hand, and guards it carefully until it can be fixed. Another robot, Sam, knocks over a glass of water onto Gillian, and apologizes profusely. Another robot, Kelly, comes across a dog whose paw is trapped in a sewer grate and is whining in pain. Kelly gently removes the paw while petting the dog's head. Another one of the robots, Frances, is taking a long walk on a snowy evening, starts shivering, and takes out some gloves and a hat and puts them on. And another robot, Ryan, is at the market purchasing cereal. Ryan stands in the aisle for a minute holding both Corn Flakes and Rice Crispies. Ryan finally puts back the Rice Crispies and places the Corn Flakes in the shopping cart.

In the case of the control vignette, participants were asked to imagine observing some humans acting in the same ways described in the final paragraph. Once participants read one of these vignettes, they were asked to answer questions according to how they understood the issues, not how they thought others might answer. The dependent measures were followed by several manipulation and comprehension checks, and then demographic questions.

4.2. Study 1: Results

Prior to the analyses testing our hypotheses, we sought to determine the internal validity of the subscales being used to measure various attributions (see Table 1). Following collection of data, coefficients of reliability were calculated for each of our 'a priori' subscales. All subscale Chronbach's alpha values were deemed to have an acceptable level of reliability (> .70).

Scale	**Corresponding Questions**	**Chronbach's Alpha**
Free Will (8 items)	These robots have free will, can make choices, have ability to do otherwise, have control over their actions, act of own free will when they act in ways we deem (im)moral, are in control (relative to programmers), and what they do is up to them.	**.842**
Moral Responsibility (5 items)	These robots are morally responsible for their actions, deserve to be blamed (relative to programmers), deserve to be punished for illegal acts, deserve to be blamed for bad acts, and deserve to be rewarded for good acts.	**.734**

Basic Emotion (9 items)	These robots can feel happiness, frustration, anger, sadness, disappointment, awe, fear, disgust, and can suffer.	**.926**
Strawsonian Emotion (9 items)	These robots can feel guilt when they do wrong, shame, pride, regret, embarrassment, admiration, indignation, and care about what happens to them and care about what happens to others.	**.918**
Conscious Sensation (4 items)	These robots *experience,* more than just *process*, the sounds in music, the images in art, the smells of food, and the softness of a blanket.	**.913**
Reason and Reflection (6 items)	These robots plan, can deliberate, can act for reasons like humans do, can have principles, can reflect on and evaluate their behavior, and can imagine alternative for future actions.	**.779**
Theory of Mind (6 items)	These robots can understand others' emotions, can empathize, can predict what others will do, can infer why others behaved, are aware of their own thoughts, and can understand their own emotional states.	**.774**

Table 1. Subscales and summarized versions of the respective individual statements (roughly one-third of these statements were worded with negations and reverse-scored). Scale validity and reliability was assessed via Chronbach's alpha.

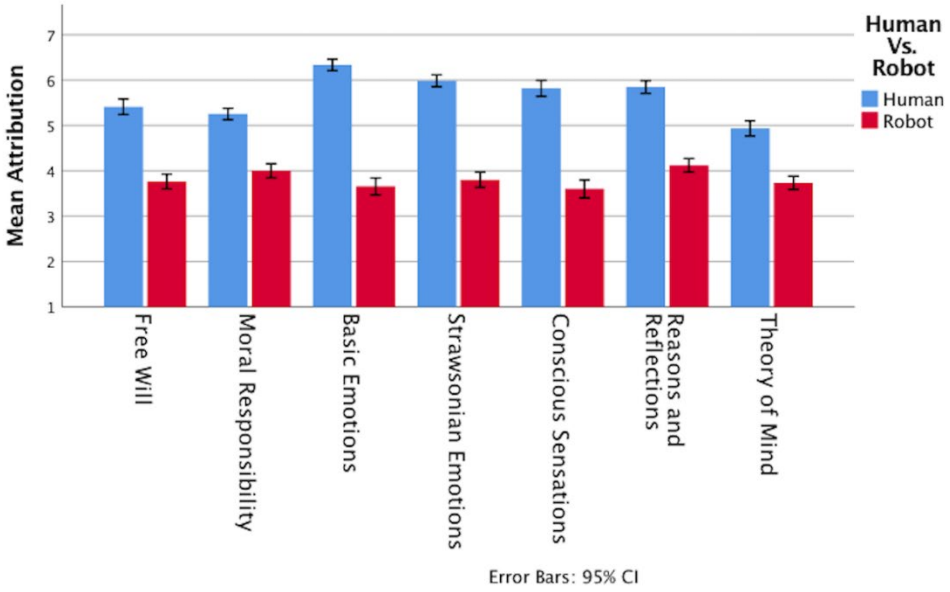
Hypothesis 1: Did attributions of conscious capacities differ between robots and humans, even though they were behaviorally indistinguishable?

Attributions of conscious capacities were subjected to a two-tailed t-test between participants within robot vignettes and those within the human control vignette. There was a significant difference in all measures of conscious capacity attribution, such that participants were less likely to attribute these conscious capacities to robots than they were to humans (see Figure 3.1). More specifically, though the robots were behaviorally indistinguishable from humans, participants attributed less free will, $t(322) = -12.62$, $p < .001$, as well as less moral responsibility to them, $t(322) = -10.72$, $p < .001$. Similarly, participants responded that these robots are less able to feel both basic and more complex Strawsonian emotions, $t(322) = -20.00$, $p < .001$, $t(322) = -17.21$, $p < .001$, respectively, as well as less able to experience sensations, $t(322) = -14.73$, $p < .001$. Participants also attributed lower levels of cognitive abilities to the robots, judging them as

being less able to reason and reflect, $t(322) = -14.94$, $p < .001$, and less able to utilize theory of mind, $t(322) = -10.28$, $p < .001$.

Note that the attributions of these capacities to robots average near the midpoint, suggesting participants were not very confident about what to say about these robots, which is unsurprising given the minimal information the vignettes provide. However, attributions became more dichotomous once we examine whether participants are considering the robots to be conscious or non-conscious (see Hypothesis 3 below).

INSERT FIGURE 3.1 AROUND HERE



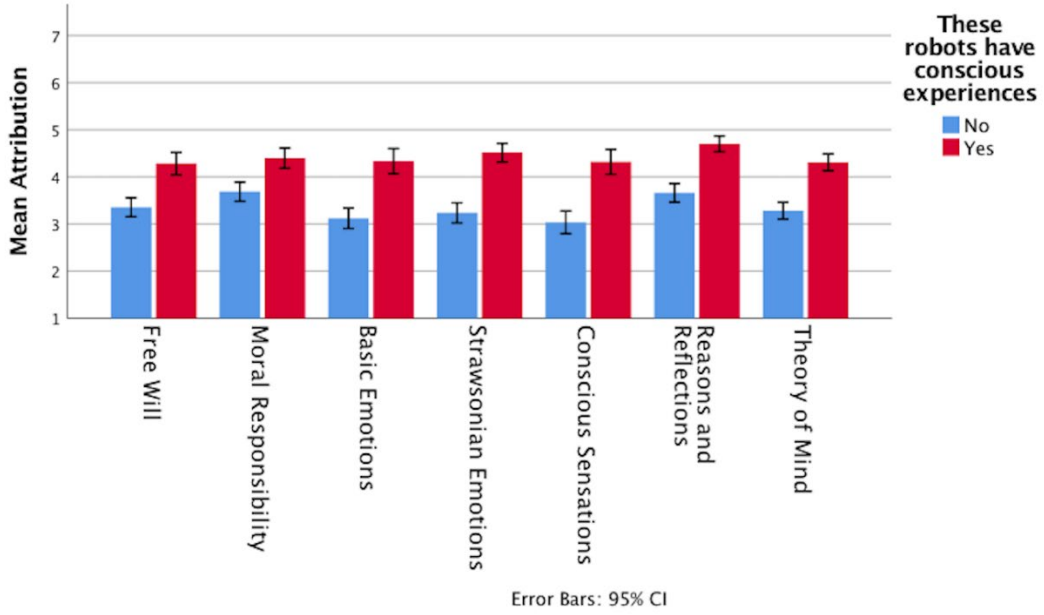
Hypothesis 2: Were participants more likely to attribute conscious capacities to robots that learned as opposed to those that were preprogrammed?

Surprisingly, the ability to learn from experience as compared to being preprogrammed had no discernible effect on any of our dependent measures (all p-values < .10 – data not shown). One possibility is that this information was less important to participants than other information about the robots. Another is that they have implicit representations of robots as fully pre-programmed that are difficult to alter with a few sentences.

Hypothesis 3: Does splitting participants by their response to the question "These robots have conscious experiences" create a divergence in the capacities that they attribute to these robots?

Upon splitting participants based on their response to a dichotomous consciousness question, we found robust differences in the capacities that were attributed to the robots (see Figure 3.2). Compared to those who responded no, those who responded that yes, these robots have conscious experiences, attributed more free will, $t(209) = -5.97$, $p < .001$, as well as more moral responsibility, $t(209) = -4.72$, $p < .001$. As expected, the same results were found for basic and Strawsonian emotions, $t(209) = -7.09$, $p < .001$, and $t(209) = -8.57$, $p < .001$, respectively, as well as the robot's ability to experience sensations, $t(209) = -7.11$, $p < .001$. Similarly, when participants saw these robots as able to have conscious experiences, they also saw them as more able to reason and reflect, $t(209) = -7.70$, $p < .001$, as well as employ theory of mind, $t(209) = -8.01$, $p < .001$. These robust results, though problematized with selection bias,[7] serve to inform the manipulations we developed for Study 2.

INSERT FIGURE 3.2 AROUND HERE



4.3. Study 2: Participants and Design

Study 2 had 474 participants (69.9% female, 30.1% male; mean age 20.0, ranging from 18-55), who were undergraduates at Georgia State University. After removing 198 for one excluded condition (see below), missed attention checks, incomplete data, or spending too little time on the survey, the final sample size was 278. Participation was voluntary and conditioned on informed consent. The study was administered online via Qualtrics.

[7] The grouping and analysis of individuals based on their responses rather than by proper randomization risks an inaccurate representation of the population originally intended to be analyzed.

This follow-up study used a 2 (Learning) x 2 (Consciousness[8]) between-groups factorial design, resulting in random assignment to all possible combinations of conditions (i.e., Learning x Conscious, Learning x Non-Conscious, Preprogrammed x Conscious, and Preprogrammed x Non-Conscious). The Learning manipulation was identical to study 1. The Consciousness manipulation varied whether the robots were described as conscious or non-conscious (as worded below). Participants' responses consisted of responses to individual statements (from 1 - "Strongly Disagree" to 7 - "Strongly Agree") attributing or not attributing different qualities to these robots, that were then combined in order to create the subscales, as described above.

The experimental vignettes were identical to experiment 1 until, following the Learning manipulation, instead of being asked to imagine the robots carrying out various specific behaviors, participants were given further information regarding the robots' mental states and capacities:

(Conscious) Furthermore, the robots are able to behave just like human beings, and they also have components that enable conscious experiences. The robots *actually feel* pain, *see* colors, and *experience* emotions. They do not *just appear* to be conscious when they carry out the same behaviors as humans.

(Non-conscious) Furthermore, the robots are able to behave just like human beings even though they do not have conscious experiences. They have components that process information such that they can carry out all the same behaviors as humans in just the same ways, but when they do so, they *just appear* to feel pain, *just appear* to see colors, and *just appear* to experience emotions.

The methods and measures were the same as those used in study 1.

4.4. Study 2: Results

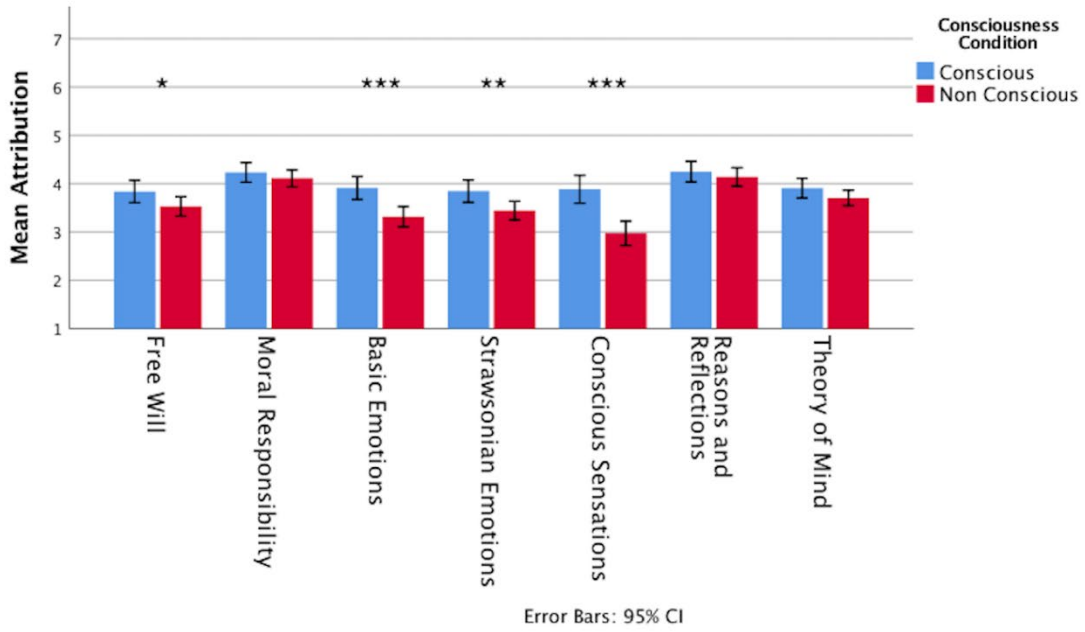
Hypothesis 1: Did attributions of capacities differ between robots described as conscious versus those that were described as non-conscious?

Attributions of capacities were subjected to a two-way analysis of variance with two Consciousness conditions (conscious vs. non-conscious) and two Learning conditions (learning vs. preprogrammed). As predicted, there was a significant main effect of consciousness on free will attributions, $F(1, 274) = 3.89$, $p = .05$, such that participants attributed less free will to robots that were described as non-conscious (see Figure 3.3). Though participants saw conscious robots as more free, they did not judge them to be more morally responsible for their actions, $F(1, 274)$

[8] For the sake of brevity, a third Epiphenomenal Consciousness condition is not included in this analysis. Responses did not differ significantly from the Consciousness condition, likely because it was difficult to get across the idea of epiphenomenalism.

= .80, $p > .05$. As expected, participants found robots described as non-conscious as less able to feel both basic and more complex Strawsonian emotions, $F(1, 274) = 13.73$, $p < .001$, $F(1, 274) = 6.89$, $p < .01$, respectively, as well as less able to experience sensations, $F(1, 274) = 22.03$, $p < .001$. However, consciousness had no discernible effect on the robots' ability to reason and reflect, $F(1, 274) = .55$, $p > .05$, nor their ability to utilize theory of mind, $F(1, 274) = 2.36$, $p > .05$.

INSERT FIGURE 3.3 AROUND HERE



Hypothesis 2: Were participants more likely to attribute conscious capacities to robots that learned as opposed to those that were preprogrammed?

As with study 1, the ability to learn from experience as compared to being preprogrammed had no discernible effect on any of our dependent measures (all p-values < .10 – data not shown).

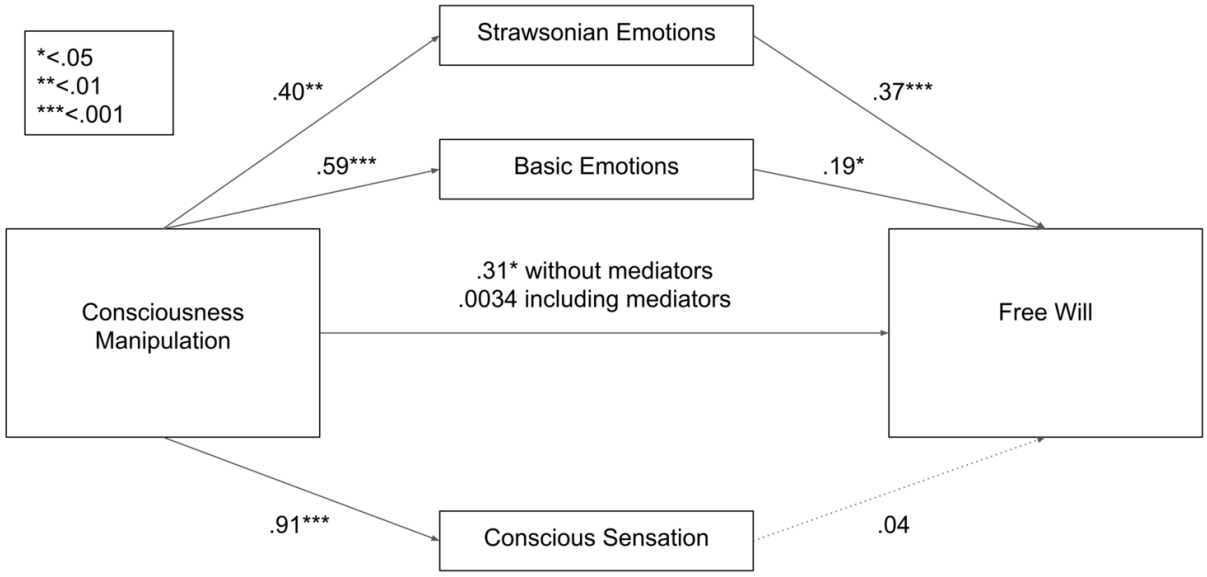
Hypothesis 3: What attributes (if any) mediated the effect of consciousness on free will attributions?

We used regression analysis in order to investigate potential mediators for the effect of consciousness on free will attributions, and selected mediators based on significant primary paths. In other words, only mediators that were directly affected by the consciousness manipulation were included in the model. Results indicate two primary mediators: participants' attribution of Strawsonian Emotions and Basic Emotions (Figure 3.4). In Step 1 of the mediation model, the regression of the consciousness manipulation on free will attribution, ignoring any

mediators, was significant, $b = .31$, $t(276) = 2.00$, $p < .05$. Step 2 showed that the regression of the consciousness manipulation on the mediators Strawsonian Emotions, Basic Emotions, and Conscious Sensations were all significant, $b = .40$, $t(276) = 2.64$, $p < .01$, $b = .59$, $t(276) = 3.72$, $p < .001$, and $b = .91$, $t(276) = 4.71$, $p < .001$, respectively. Step 3 of the mediation analysis showed that, while controlling for the consciousness manipulation, the emotional mediators (Strawsonian and Basic) were each significant predictors of Free Will attribution, $b = .37$, $t(273) = 4.06$, $p < .001$ and $b = .19$, $t(273) = 2.05$, $p < .05$, while Conscious Sensation was not, $b = .04$, $t(273) = .79$, $p = .43$. Step 4 of the mediation analysis revealed that, while controlling for Strawsonian and Basic Emotions (as well the negative control of Consciousness Sensation), the consciousness manipulation was no longer a significant predictor of Free Will attribution, $b = .0034$, $t(273) = .026$, $p = .98$, 95% CI [-.26, .26], indicating full mediation.[9] Thus, it was found that the extent to which people judged the robots as able to experience Strawsonian and Basic Emotions fully mediated the relationship between the consciousness manipulation and people's attributions of Free Will.

These results suggest that phenomenal consciousness plays a particular role in the attribution of Free Will, but not an indiscriminate role. In other words, the ability to feel emotions, which suggests that outcomes actually matter to the individual, is important in Free Will attributions, yet the ability to have conscious sensations specifically (e.g., the ability to *experience* sounds or smells) plays no significant role.

INSERT FIGURE 3.4 AROUND HERE



[9] The total effect was tested using a bootstrap estimation approach within Andrew Hayes' PROCESS with 5000 samples (2012).

5. Conclusions

Our results provide some support for our proposed connection between consciousness and free will. However, further studies are clearly required to allow a more fine-grained understanding of which features of consciousness matter most to people's attributions of free will, as well as their relation to attributions of moral responsibility, where we found inconsistent results, and also to determine what roles learning and experience play in these attributions, given that our manipulations of whether the robots learn or are fully pre-programmed did not have significant effects. If future research bolsters our initial findings, then it would appear that when people consider whether agents are free and responsible, they are considering whether the agents have capacities to feel emotions more than whether they have conscious sensations or even capacities to deliberate or reason. It's difficult to know whether people assume that phenomenal consciousness is required for or enhances capacities to deliberate and reason. And of course, we do not deny that cognitive capacities for self-reflection, imagination, and reasoning are crucial for free and responsible agency (see, e.g., Nahmias 2018). For instance, once considering agents that are assumed to have phenomenal consciousness, such as humans, it is likely that people's attributions of free will and responsibility decrease in response to information that an agent has severely diminished reasoning capacities. But people seem to have intuitions that support the idea that an essential condition for free will is the capacity to experience conscious emotions. And we find it plausible that these intuitions indicate that people take it to be essential to being a free agent that one can feel the emotions involved in reactive attitudes and in genuinely caring about one's choices and their outcomes. If so, these intuitions support the sort of self-expressive views built on the foundations laid by Strawson and Frankfurt.

We do not want to defend here a metaphilosophical account of the role of ordinary intuitions in philosophical theorizing, a topic of much recent controversy. We will simply conclude by pointing out that most people, along with most theorists, seem to think that consciousness is crucial for free will. Few theorists offer adequate explanations of the connection. If a theorist aims to reject the importance of consciousness to free will, she should explain both what drives most people to think otherwise and why those people are mistaken. If a theorist aims to understand the connection, it might help to understand why ordinary people see it. In fact, we think that understanding why philosophers and non-philosophers alike think that there is a connection between consciousness and free will might suggest strategies for developing plausible theories that explain the connection. If our above results are any indication, these theories will focus on our capacities to *actually care* how others treat us and how we treat them, to *feel* reactive attitudes in response to such treatment, and to *experience* the emotions necessary for caring about our decisions and the outcomes of those decisions.

Perhaps, fiction points us towards the truth here. In most fictional portrayals of artificial intelligence and robots (such as *Blade Runner*, *A.I.*, and *Westworld*), viewers tend to think of the robots differently when they are portrayed in a way that suggests they express and feel emotions. No matter how intelligent or complex their behavior, the robots do not come across as free and autonomous until they seem to *care* about what happens to them (and perhaps others). Often this

is portrayed by their showing fear of their own or others' deaths, or expressing love, anger, or joy. Sometimes it is portrayed by the robots' expressing reactive attitudes, such as indignation about how humans treat them, or our feeling such attitudes towards them, for instance when they harm humans. Perhaps the authors of these works recognize that the robots, and their stories, become most interesting when they seem to have free will, and that people will see the robots as free when they start to care about what happens to them, when things really matter to them, which results from their consciously experiencing the actual (and potential) outcomes of their decisions and actions.

References

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Baumeister, R. (2008). Free will, consciousness, and cultural animals. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will*, (pp. 65–85), Oxford University Press.

Bargh, J. A. (2008). Free will is un-natural. In J. Baer, J. C. Kaufman & R.F. Baumeister (Eds.), *Are We Free?: Psychology and Free Will,* (pp. 128-154), Oxford: Oxford University Press.

Berlin, I. (1958). Two concepts of liberty. In *Four Essays on Liberty.* Oxford: Oxford University Press.

Block, N. (1995). On a confusion about the function of consciousness. *Behavioral and Brain Sciences*, 18, 227–247.

Buss, S. (2012). Autonomous action: Self-determination in the passive mode. *Ethics*, 122, 647–691.

Caruso, G.D. (2016). Consciousness, free will, and moral responsibility. In R.J. Gennaro (Ed.), *The Routledge Handbook of Consciousness*, (pp. 78-90), New York: Routledge.

Clarke, R. (2003). *Libertarian accounts of free will*. New York: Oxford University Press.

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

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

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

Frankfurt, H.G. (1987). Identification and wholeheartedness. In F.D. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, Cambridge: Cambridge University Press.

Frankfurt, H.G. (1999). *Necessity, volition, and love*. Cambridge: Cambridge University Press.

Harris, S. (2012). *Free will*. New York: Free Press.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf

Hodgson, D. (2002). Quantum physics, consciousness, and free will. In R. Kane (Ed.), *The Oxford Handbook of Free Will*, (pp. 57-83), Oxford: Oxford University Press.

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

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?. *Phenomenology and the cognitive sciences*, *9*(1), 133-155.

Jack, A. I., & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, *3*(3), 383-403.

Kane, Robert (1998). *The Significance of Free Will*. Oxford: Oxford University Press.
Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the cognitive sciences*, *7*(1), 67-83.

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. (2002). Do we have free will? In R. Kane (Ed.) *The Oxford Handbook of Free Will*, (pp. 551-564), New York: Oxford University Press (Reprinted from *Journal of Consciousness Studies,* 1999, 6, 47-57).

Mele, A.R. (2010). Conscious Deciding and the Science of Free Will. In R.F. Baumeister, A.R. Mele, & K.D. Vohs (Eds.), *Free Will and Consciousness: How Might They Work?*, (pp. 43-65), New York: Oxford University Press.

Monroe, A.E., Dillon, K.D., & Malle, B.F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition,* 27, 100-108.

Monroe, A. & Malle, B.F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology*, 1, 211–224.

Murray, D. & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*, 88, 434–467.

Nahmias, E. (2018). Free will as a psychological accomplishment. In D. Schmidt, & C. Pavel (Eds.), *The Oxford Handbook of Freedom* (pp. 492-507), New York: Oxford University Press.

Nahmias, E. (2014). Is Free Will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Eds.), *Moral Psychology, vol. 4, Free Will and Moral Responsibility*, (pp. 1-25), New York: MIT Press.

Nahmias, E., Morris, S.G., Nadelhoffer, T., & Turner, J. (2006) Is incompatibilism intuitive? *Philosophy and Phenomenological Research,* 73(1), 28-53.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41, 663–685.

Penrose, R. (1991). The emperor's new mind. *RSA Journal*, 139, 506-514.

Sartre, J.P. (1943 [2003]). *Being and nothingness: An essay on phenomenological ontology*. London: Routledge.

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, 929-946.

Shepherd, J., & Levy, N. (forthcoming) Consciousness and morality. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness*, Oxford University Press.

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

Shoemaker, D. (2003). Caring, identification, and agency. *Ethics,* 114, 88-118.

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

Sripada, C. (2016). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies*, 173, 1203-1232.

Stapp, H. P. (2001). Quantum theory and the role of mind in nature. *Foundations of Physics*, 31(10), 1465-1499.

Stillman, T.F., Baumeister, R.F., & Mele, A.R. (2011). Free will in everyday life: Autobiographical accounts of free and unfree actions. *Philosophical Psychology*, 24, 381-394.

Swinburne, R. (2013). *Mind, brain, and free will*. Oxford: Oxford University Press.

Strawson, G. (1986). *Freedom and belief*. Oxford: Oxford University Press [revised edition 2010].

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

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

Vonasch, A., Baumeister, R., & Mele, A., (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. *Consciousness and Cognition*, 60, 133-151.

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

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

Wolf, S. (1990). *Freedom within 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.