Responsibility and Consciousness

Matt King and Peter Carruthers

In D.K. Nelkin and D. Pereboom (Eds.), *Oxford Handbook of Moral Responsibility*. Oxford University Press (2022), 448-467.

1. Introduction

Intuitively, consciousness matters for responsibility. A lack of awareness generally provides the basis for an excuse or at least for blameworthiness to be mitigated. If you are aware that what you are doing will unjustifiably harm someone, it seems you are more blameworthy for doing so than if you harm them without awareness. There is thus a strong presumption that consciousness is important for responsibility. The position we stake out below, however, is that consciousness, while relevant to moral responsibility, is not necessary.

The background for our discussion is an emerging consensus in the cognitive sciences that a significant portion, perhaps even a substantial majority, of our mental lives takes place unconsciously. For example, routine and habitual actions are generally guided by the so-called "dorsal stream" of the visual system, whose outputs are inaccessible to consciousness (Milner & Goodale 1995; Goodale 2014). And there has been extensive investigation of the processes that accompany conscious as opposed to unconscious forms of perception (Dehaene 2014). While there is room for disagreement at the margins, there is little doubt that our actions are much more influenced by unconscious factors than might intuitively seem to be the case. At a minimum, therefore, theories of responsibility that ignore the role of unconscious factors supported by the empirical data proceed at their own peril (King & Carruthers 2012). The crucial area of inquiry for those interested in the relationship between consciousness and responsibility concerns the relative strength of that relationship and the extent to which it should be impacted by findings in the empirical sciences.

While the nature of consciousness continues to be much debated, most of those debates concern the nature and extent of so-called "phenomenal" consciousness (which is how experiences *feel*, or what they are *like* for their subjects). This is arguably irrelevant to our topic, for the intuition that consciousness is tightly connected to responsibility has little to do with the "feely" character of conscious experience, we

submit, and much more to do with *awareness*.[1] It is lack of awareness of the circumstances or consequences of one's actions that is thought to excuse, not the presence or absence of felt qualities of one's experiences while acting. And there is much less dispute about the nature of awareness, which seems to coincide with the notion of so-called "access consciousness" (Block 1995). Moreover, the best—and widely accepted—theory of access consciousness is the "global workspace" or "global broadcasting" model (Baars 1988, 2002; Dehaene 2014). On this account, mental representations become conscious when they pass a threshold of activation (generally resulting from attentional signals being directed toward them), which results in them being made widely available to multiple subsystems in the brain—specifically, to systems for planning, decision-making, and verbal report; for forming long-term memories; and for issuing in full-blown emotional reactions. At any rate, this is the framework for thinking about consciousness that we adopt in the present chapter.

2. The relevance of consciousness

Even if consciousness is intuitively relevant to responsibility, it remains an open question precisely how and to what extent it is relevant. One natural suggestion is that consciousness is in some sense *required* for responsibility. In order to be responsible for an action, its main psychological causes (not only its goals, but the beliefs and perceptions that guide it) must be conscious. Call this the *consciousness condition*. While we agree that consciousness is relevant to responsibility, we deny the consciousness condition.

The consciousness condition has some intuitive support. We often attempt to excuse our behavior by pointing out that we were unaware of relevant details. It is customary to excuse bumping into someone by saying "I didn't see you there." When we lack important information, our ignorance can serve to mitigate blame. More significantly, we are more apt to excuse behavior the more unconscious it is. In some cases, such as that of a sleepwalker, we may even exculpate the agent altogether.

Notably, however, it is just as apparent that we don't *always* excuse people for lack of awareness. If I forget a friend's birthday, I'm apt to feel guilty about doing so and thus apologize. Moreover, such responses appear rational. We hold others to account not just for what they notice and do, but what they fail to notice, and so fail to do (Smith 2005). It is common practice to blame people for their negligent actions, where they are unaware of an unjustified risk of harm their conduct poses (usually through a failure

[1] Though see Shepherd 2015. See also n.8.

to take reasonable care).[2]

There is disagreement about whether such cases really show that awareness isn't required. Defenders of the consciousness condition might insist that responsibility is only possible in such cases through some prior exercise of consciousness-involving responsibility. So, while the agent at the time is unaware that today is her friend's birthday, we can nonetheless "trace" her responsibility back to some prior responsible action (or omission).[3] Such claims are controversial, however. First, any such tracing often leaves us with an identical problem to resolve. For example, if one traces culpability for forgetting a friend's birthday back to culpability for failing to take sufficient steps to remember (such as putting a reminder in one's calendar), then we would still have to explain responsibility for that prior forgetting.[4] So we would need some final terminus for responsibility wherein the consciousness condition was satisfied. That seems implausible for a wide variety of cases. Second, one's culpability for some prior act (or omission) is often significantly less than one's culpability for the present one, so the strategy still leaves us without a fully adequate explanation.[5]

While we allow that *some* cases of unconscious action might be explained via tracing back to some previous conscious episode, we deny that all cases are best explained this way. A full defense of this claim, however, would steer us too far aside from the main discussion. Instead, we will simply observe that we often fail to excuse others for lack of awareness, and in doing so we do not always look to trace their culpability back to some prior episode.

More dramatically, not only does the fact that some action was unconsciously-caused not always mitigate responsibility, it occasionally seems to exacerbate it. It is a familiar thought that what we do spontaneously and without conscious reflection can reveal a deeper truth about ourselves than more careful, deliberate action.

Rather than settling matters, then, reflection on our ordinary practices reveals that we have diverging commitments when it comes to the consciousness condition. On the one hand, we often excuse behavior precisely because of its unconscious nature. On the other hand, we don't universally excuse such

---

[2] This is not to say such practices are without controversy. See King 2009.

[3] E.g., see Smith 1983; Wieland & Robichaud 2017. For alternative strategies, see King 2014, 2017; Sher 2009; Smith 2011.

[4] See King 2009, 2014 and Sher 2009 for critiques of this form.

[5] To use an example from Sher 2009, forgetting one's dog in the car on a hot day is worse than deciding to bring the dog with you while running errands on a hot day. And, of course, one would still need to show that one was conscious of the risk to the dog for the tracing strategy to work.

behavior. Any discussion of whether, and in what sense, consciousness is necessary for responsibility will need to engage with these observations.

Beyond our ordinary experience, scientific findings raise a variety of concerns. Some results have been taken to suggest that consciousness plays no significant role in the production of action or responsible choice (Libet 1993; Wegner 2002). This would suggest, if consciousness is required for responsibility, that we are much less responsible than we think. But such research has been widely undermined both philosophically and empirically.[6] This shouldn't be surprising, since it is quite implausible to suppose that consciousness has *no* role to play in action-production and responsibility, even if it isn't strictly necessary.

More plausible worries concern the manner in which our actions and decisions might be affected or directed in ways of which we are unaware. For instance, it has been repeatedly found that people can have affective biases of which they lack awareness (and which influence their everyday behavior), as well as having their expectations of other people influenced by racial, gender, and other stereotypes that they wouldn't endorse consciously (Banaji & Greenwald 2013).[7] There is plenty of disagreement about the significance of such findings. Controversy over the research is perhaps to be expected given our general ambivalence about the precise relevance of consciousness to responsibility. As indicated above, unconscious factors seem to excuse in some kinds of cases but not in others, while in still others they may actually heighten responsibility. Given the lack of consensus, both from our ordinary practices and scientific research, greater attention is required to refine the relationship between consciousness and responsibility.

In the limited space available to us, it is impossible to canvass every possible variant of the consciousness condition, nor every area in which consciousness (or unconsciousness) might be implicated. In light of this, we plan to examine the consciousness condition through three of the types of case so far mentioned: automaticism (like sleepwalking), forgetting, and implicit bias. While by no means exhaustive, they represent a suitable spectrum along which to pursue the issue.

---

[6] For a thorough review of the original research and subsequent critiques, see Levy 2014, pp. 14-26. Additionally, it's worth noting that most appeals to Libet- and Wegner-style experiments *assume* something like the consciousness condition, rather than arguing to it from the data.

[7] In addition to implicit bias, some claim that seemingly irrelevant features of our situation can affect our judgments or reasons-responsiveness in ways of which we aren't aware (i.e., the so-called "situationism" literature). To cite just one example, having previously held a hot cup of coffee can positively influence one's assessment of others (Williams & Bargh 2008). But there is simply no consensus either on the exact nature of these effects, their scope, or the degree to which they threaten responsibility. Some relevant recent discussions include Caruso 2015, Herdova 2016, Holroyd 2012, Mavda 2018, and Schlosser 2013.

3. Some generalities regarding responsibility

Before proceeding to cases, however, it is important to consider in what way consciousness might be relevant to responsibility generally. After all, while consciousness could be brutely necessary, in such a way that the consciousness condition would be independent of any other necessary conditions on responsibility, such a position would be explanatorily disappointing.[8] We would be left without an answer for *why* consciousness matters. Instead, it is more promising to suppose that if consciousness is necessary it is because it makes possible or facilitates the satisfaction of some other condition on responsibility. Indeed, even though we sometimes excuse people for actions done in ignorance, we don't generally think this is *simply* because they lacked awareness. Rather, a more plausible idea is that consciousness tends to correlate with, or is indicative of, something else that is itself essential for responsibility. At least, this is how we propose to proceed. We will examine the prospects for the consciousness condition within a basic framework of what *else* matters for responsibility.

There is a plethora of theories of moral responsibility. Indeed, it is fashionable these days to distinguish different conceptions of responsibility, which differ from one another in their dimensions or requirements.[9] For our purposes, we propose to work with a unified, monistic conception of responsibility, in which to be responsible is to deserve (or to merit) blame for bad things and praise for good things. This is, in our estimation, a perfectly common conception and is the one most often invoked in traditional philosophical debates about free will. Moreover, it arguably comes with the most stringent conditions.[10] So if consciousness isn't required for responsibility thus conceptualized, we should expect the result to

[8] Shepherd 2015 reports that folk judgments of responsibility do seem to be influenced by judgments about phenomenal consciousness. But it isn't clear from the findings what the precise relation between those two judgments is or how to apply them to informational awareness. In any case, even if the folk judged consciousness to be necessary, that wouldn't show it was directly so.

[9] David Shoemaker (2015) has perhaps the strongest such defense, contending that there are multiple independent conceptions of responsibility that differ in terms of their conditions and associated warranted responses. Others contend, for example, that there are conceptions that do not involve any claims about desert (Pereboom 2016). While growing in popularity, this approach is not without its opponents (cf. Smith 2012).

[10] Among those that are remotely plausible, that is. We grant that ours isn't the *most* demanding conception out there. Strawson 1986, for instance, characterizes moral responsibility to be a matter of one's deserving *eternal* punishment or reward. That strikes us (and many others) as far too demanding a notion, and one that doesn't correspond to whatever conceptions might be at play within our ordinary practices of holding others responsible.

generalize to other accounts as well. Nevertheless, we don't rule out that there might be further complexities that arise in evaluating the prospects for the consciousness condition across other plausible conceptions of responsibility. Instead, we aim for a useful generality in our discussion, so that those who prefer alternative characterizations of responsibility can be invited to modify our discussion accordingly. We believe the basic claims will remain unchanged.

Indeed, no matter the account one offers of responsibility, some general features of agency tend to be prominent: specifically choice, control, and coherence.[11] These features are not exclusive of one another, and there is plenty of disagreement over which are necessary for or most central to responsibility. Nevertheless, they represent a plausible core of responsibility-relevant capacities.

Consider choice first. Some accounts state that responsibility requires agents to select from among genuinely available alternatives, free of external constraints (Kane 1996). What matters here is the agent being free to choose. (While not required, most such views favor incompatiblism between free will and determinism.)

Other theories stress the capacity to exercise effective control (Mele 1995). Since actions that an agent didn't control, like accidents, look to be one's for which she isn't responsible, a popular requirement on responsibility is that the agent guided her conduct appropriately. Many theories develop this idea in terms of being "reasons responsive" (Fischer & Ravizza 1998). Agents exercise the relevant control in cases where they are capable of being sensitive to the salient moral features of their circumstances, and of guiding their behavior appropriately in light of those reasons.

Still other theories emphasize the coherence between the action (and its motivations) and the agent's overall psychology (Frankfurt 1971; Sripada 2016). The thought is that an agent is more responsible for an action the more it reflects her moral self, or the better integrated its motivations are with the full set of her motivations and values. Consider the difference between a committed racist who utters a racist remark and someone who simply makes an insensitive comment (one that is out-of-character). There is an intuitive difference here that is captured by the idea that responsible action is one that manifests the agent's true commitments and values.

We can calibrate our understanding of the consciousness condition to these core capacities. On the first idea, if what matters is choice, we might wonder whether consciousness is required for choices to be genuinely available or properly selected by an agent. On the second, we might wonder whether we can

[11] The following characterizations are modeled on King & May 2018.

respond in the requisite way to reasons represented only unconsciously. On the third, we can wonder whether unconscious mental processes properly reflect our values or not.

Let us first consider, and quickly reject, the idea that choice among alternatives grounds the consciousness condition. While few claim that consciousness is required for free choice, the idea is not without its defenders. The thought is that consciousness provides the space within which we might find indeterministic mental processes, sufficient to underwrite the sorts of distinctive capacities that free will and responsibility require (Hodgson 2012). If neural processes are governed by deterministic physical laws, perhaps consciousness might supply the right sort of non-deterministic arena in which free choices can occur.

Such views are not popular, and for good reason. First, they are overly speculative, often optimistically relying on future science to vindicate their claims. Second, they confront a dilemma, depending on whether or not physicalism about consciousness is endorsed. If physicalism is maintained, then it would have to be claimed that only certain sorts of neural processes are indeterministic (namely, those implicated in consciousness), because otherwise it would be possible for the requisite indeterminism to be present in unconscious processes. If physicalism is denied, on the other hand, then it would need to be claimed that nonphysical processes (e.g. conscious deliberation) can have physical effects (action). The first horn of the dilemma leaves us with two fundamentally different types of neural processes, whereas the second conflicts with the causal closure of physics.

An alternative way of motivating the consciousness condition on choice might be to claim that the ultimate foundation for responsibility requires a kind of deliberative competition between options. For example, Kane (1996) argues that we are responsible for our choices because of previous "self-forming actions," which are (roughly) decisions in which an agent is torn in two directions (e.g., whether to take a new job in a new city or stay where one is), and it is undetermined which option they will choose. Whatever the agent decides will, in some sense, help shape the person they are to be. One might then insist that such indeterministic competition only occurs in conscious deliberation.

However, we still don't have good grounds for the consciousness condition. We have already seen that it is unlikely that consciousness can supply the requisite indeterminism. And in addition Kane must claim that the kind of choice that leads to self-forming actions will always involve conscious deliberation. This means limiting responsible actions to those that are "self-forming" in the required sense (or that flow from aspects of one's character fixed by such actions). While we cannot definitely reject such a view about

responsibility here, it is a significant, and in our view implausible, commitment to take on.[12]

Although we leave aside the relevance of consciousness to choice, it might still be relevant for control or coherence. Indeed, the most recent and sustained defense of the consciousness condition argues that consciousness is required on either of these two grounds. This is where we go next.

4. Defending the consciousness condition

In the most thorough defense to date, Neil Levy (2014) argues that what matters for responsibility is *informational awareness*. It is our lack of awareness of morally relevant features of our conduct that excuses us, when we are excused. Drawing on the global-broadcasting model of consciousness, he holds that the basic function of consciousness is to integrate our attitudes with one another and coordinate our activities over time. As such, consciousness is necessary for a kind of rational agency. Behavior driven by conscious states will be better integrated—more coherent and consistent—than behavior driven by unconscious states. Consciousness is thus required to control our behavior properly and to display the requisite flexibility in our responsiveness to reasons.

Additionally, Levy argues that consciousness is necessary if a coherence-condition is to be fully met. Since conscious contents are those that are globally broadcast and made available to other consuming systems in the mind, the contents of beliefs, values, and other attitudes, when consciously entertained, will be better integrated with our other beliefs, values, and attitudes. Only conscious contents can be compared with and evaluated against the full suite of our other attitudes, beliefs, and values. Thus, when we are conscious of the springs of our actions those actions more fully express our true selves, since more of our evaluative stance has been brought to bear on them. The conscious decisions we come to, and the actions we then perform, result from competition among a larger (perhaps maximal) set of our personal-level attitudes.

Levy emphasizes personal-level attitudes because, he says, only mental representations that are broadcast widely in the mind will be appropriately predicated of the person herself, rather than some lower-level component of her mind. Since we hold *persons* morally responsible, we are interested only in facts about the person rather than about her subsystems. I am no more responsible for how certain components of my visual system operate than I am for my digestion, so the thought goes. And in order for

[12] Notice, also, that such a view will be required to take on some sort of "tracing" mechanism, to account for the very plausible cases of responsible choice that do not involve torn decisions.

me to be informationally aware of something, it must be personally available—that is, available as a personal-level representation.

On Levy's account, *personally-available information* is information that is (1) easily and effortlessly retrievable and (2) *online*. Information that is easily retrievable is apt to be triggered by many ordinary cues (rather than leading questions, say). To be online is to actually play a role in guiding behavior. As an empirical fact, according to Levy, the only information that will meet both conditions is conscious information (that is, information that is globally broadcast).

Some have denied that consciousness, in this sense, is necessary for responsibility. For instance, Arpaly (2002) notes a number of cases in which agents act and are seemingly responsible for doing so despite their motivations remaining introspectively opaque. They are not aware, it seems, of why they do the things they do, and yet they strike many of us as responsible, even praiseworthy, for acting. Huck Finn is a popular case study. In a famous scene of Mark Twain's novel, Huck has an opportunity to turn in the escaped slave Jim to a group of men hunting for him. Huck believes he ought to reveal Jim's location (for he shares the mistaken moral views of his time), but he can't bring himself to do it, whether out of unconscious sympathy or an unrecognized recognition of their shared humanity. Interpretations of the case vary, but at least some theorists are led to deny a consciousness condition because of such cases, wherein agents act for reasons of which they are unaware and yet reveal their praiseworthy or blameworthy qualities. Likewise, we at least occasionally experience guilt when we realize that our past motivations for some action were less-than-noble. And we are all familiar with cases where people have "blindspots". In such cases we recognize the true motivations of another (and hold her responsible accordingly), although the person herself is unaware of those motives.

Levy is committed to denying responsibility in such cases. Though agents may sometimes have their actions guided by online representations, this is insufficient for responsibility if those guiding representations fail to be incorporated into their personal-level representations. He gives the example of a woman who forgets her anniversary and so plans a date with friends, but in getting dressed she chooses a particular necklace to wear *because* it is the necklace her spouse gave her on their first anniversary. The belief that today is her anniversary is online, since it is guiding her behavior, but while she might be able to retrieve that information if asked a leading question, it isn't easily retrievable under a wide variety of circumstances. This seems to follow as a consequence of her forgetting the anniversary in the first place. For she is likely to have encountered cues recently that should have led to retrieval of the belief. For example, she saw her husband multiple times over the past week, she looked at the date when planning the dinner

with friends, and she regularly passes photos from their wedding in the hallway. If she forgot their anniversary despite regular cues related to that belief, it would seem that "the mere fact that some of her behavior is guided by the knowledge that it is her wedding anniversary does not seem to establish that [she is responsible for forgetting]" (33).

Information that is only easily retrievable is likewise insufficient to underpin responsibility, for Levy. This is because one can, through happenstance, fail to encounter any of the cues that would lead to easy retrieval of the information if they were present. Suppose, for example, that had the woman been home she would have remembered her anniversary as she always does, but she is away on a trip in the days leading up to it, travelling in a remote part of the country with no contact with her husband. With nothing to remind her, she may remain unaware of the relevant information, and we would be more apt to excuse in these circumstances. As Levy says, the information needs to be both easily retrievable *and* online for responsibility.

For Levy, then, consciousness is generally relevant because it affords us the ability to integrate divergent beliefs and values, and to coordinate our activities over time. More specifically, only when we are conscious of the morally significant features of our actions is such information personally available and hence reflective of the person's moral agency. Only then can we be responsible for our actions.

5. Unconscious evaluation

Levy's defense of the consciousness condition is unconvincing, and its shortcomings suggest we should be skeptical of any such necessary condition on responsibility. In this section, we argue that the relationship between consciousness and responsibility is more complicated than the consciousness condition allows. We agree that consciousness is *relevant* to responsibility, but we are responsible for much that is unconscious. To bring out these points, we return to three common (and divisive) kinds of case in the literature: automaticism (e.g., sleepwalking), forgetting, and implicit bias.

Automaticism

It can be tempting to think there must be some necessary consciousness condition. After all, at the extreme, there are cases of apparent automaticism, where even relatively complicated actions are performed by individuals who nonetheless have little or no awareness of what they are doing. Kenneth Parks provides a vivid (and disturbing) example.

In May of 1987, after falling asleep, Parks got up, got dressed, and drove 23 kilometers to his in-laws' house, whereupon he assaulted his father-in-law and repeatedly stabbed his mother-in-law, before driving to the police station. Once there, he told police that he thought he had killed people, only then (apparently) noticing the severe wounds to his own hands (Broughton et al. 1994). Parks' memory of the events was patchy at best, and his defense of somnambulism was credible given a history of disordered sleep, plausible triggering conditions (he was under extreme stress at the time), and lack of any motive (he got along very well with his in-laws). The jury agreed, and Parks' acquittal under a plea of "non-insane automatism" was upheld by the Canadian Supreme Court (*R. v. Parks* 1992).

The evidence suggests that most of what Parks did was driven by low-level, unconscious representations. And it might be on those grounds alone that he was acquitted. But we think it would be a mistake to conclude from this that consciousness is necessary for responsibility. This is because the underlying causes of sleepwalking are not equivalent to the full range of unconscious processes that operate in normal non-pathological behavior. Two findings are especially relevant here. The first is that sleepwalking occurs during the deepest phases of non-rapid-eye-movement sleep (Plazzi et al. 2005). The second is that sleepwalking is associated with activation of sensory-motor pathways between the thalamus and motor regions of cingulate cortex, combined with sustained de-activation of most other regions of cortex, including prefrontal executive and decision-making areas (Bassetti et al. 2000; Terzaghi et al. 2012; Januszko et al. 2016). Moreover, although the amygdala can remain responsive to low-level threat-like stimuli during sleepwalking, the remaining components of the sleepwalker's cortical and subcortical evaluative networks are likewise de-activated (Terzaghi et al. 2012). A plausible explanation for why Parks wasn't responsible for his actions, then, is not merely that his actions were unconsciously caused, but rather that they were caused in ways that bypassed both his value systems and his decision-making capacities, all of which were shut down by sleep. His actions thus reflected nothing of his beliefs and values.

Forgetting

The fact that a process is unconscious does not by any means entail that it bypasses relevant evaluative systems of the agent. This is significant, for if unconsciously-caused action can still reflect an agent's assessment of her reasons or values, then one could still be responsible for unconscious action, either through control or coherence. Indeed, there are grounds for thinking that evaluation is already implicated in determining whether or not a content becomes conscious in the first place. In particular, the evidence suggests that whether or not some information gets globally broadcast often reflects a competition for

attention, resulting in a decision to attend (albeit an unconscious one). This claim will require some background and elucidation.[13]

It is widely agreed among cognitive scientists that working memory provides the workspace within which representations can be consciously activated, sustained, and manipulated (Miyake & Shah 1999; Baddeley 2003; Alloway & Alloway 2013). It is also widely agreed that entry into working memory (and consciousness) depends heavily on the direction of top-down attentional signals, which are under intentional control (Gazzaley 2011; Tamber-Rosenau et al. 2011; Unsworth et al. 2015). These are thought to boost the activity of some representations over the threshold for global broadcasting while at the same time suppressing others. Hence there are such phenomena as inattentional blindness, where people can remain unconscious of even quite salient stimuli because their attention is directed elsewhere (Dehaene 2014).[14]

In addition to the top-down attentional network controlled by current goals and intentions, there is also a competing network—generally described as a *saliency* network—whose function is to compete for the resources of top-down attention, allowing a new set of contents to become conscious (Corbetta & Shulman 2002; Corbetta et al. 2008; Menon & Uddin 2010). The term "saliency" isn't really appropriate to describe this network, however, since it is sensitive to much more than just low-level features such as stimulus intensity (a loud bang) or sudden change (a shape starting to wriggle in the grass as one approaches). On the contrary, the network also evaluates unattended stimuli (as well as unattended memories that may have become active in the context) against standing values and standing goals (Awh et al. 2012). So it is better described as the *relevance* network.

The interactions between these two systems are responsible for the well-known cocktail-party effect. One might be fully engaged in conversation with someone at a party, attending to (and thus conscious of) what they are saying. But at the same time the relevance network monitors the surrounding conversations. If someone happens to use one's name, this is apt to win the competition for top-down attentional resources, and the sound of the name pops suddenly into one's stream of consciousness. One

[13] The ideas outlined in the next few paragraphs are fully developed, and supported with an extensive survey of the empirical literature, in Carruthers 2015.

[14] In one famous example, participants watching a basketball game are directed to count the number of times a player in a white t-shirt touches the ball. In the course of the game a man in a gorilla suit walks onto the center of the court, beats his chest, and walks off. Participants tasked with attending to the players in white t-shirts often fail to notice the gorilla altogether, despite his being fully visible to participants *not* so tasked. Examples are readily available on YouTube.

has, in effect, *decided* that the mention of one's name is more relevant to one's concerns than whatever one's friend happens to be saying, and so one directs one's attention accordingly. Indeed, scientists who model such phenomena think that one is engaging in the functional equivalent of a cost-benefit analysis, weighing the costs and benefits of one's current focus of attention against the alternatives (Kurzban et al. 2013).[15]

The interactions between these networks are also thought to be responsible for the phenomenon of mind-wandering (Fox et al. 2015). One might be attending to a task, such as reading an article for work. But at the same time the relevance network will be monitoring surrounding stimuli as well as memories that have been sparked into activity by the circumstances. If any of these surrounding stimuli or memories are deemed relevant enough (and especially if one's primary task is experienced as aversive), then one will be apt to find oneself suddenly imagining a tropical beach or planning a recipe for dinner.

With this background in place, return to the phenomenon of forgetting. Consider Levy's example of the wife who forgets that today is her anniversary. We surmised earlier that any number of cues *could* have activated the belief, from passing some wedding photos to looking at the date on the calendar when planning her meeting with friends. Levy concludes that the belief isn't personally available because it wasn't in fact activated and globally broadcast. An alternative possibility, however, is that it wasn't broadcast precisely because it wasn't judged important enough given her current goals and values. The sight of the date might indeed have evoked the knowledge that it is her anniversary, but this wasn't deemed relevant enough to attract attention, given her focus on arranging her meeting. As a result, it never became conscious.

Recall that Levy denies that any (unconscious) component of the mind can constitute a state or attitude that is attributable to the person herself, rather than just a component. Only consciously entertained representations are properly *of* the person (personal-level), as they better reflect the agent's evaluative stance, having been available for competition and coordination with the agent's other values and goals. Only such representations can thus ground responsibility for subsequent action. However, rather than failing to be a personal-level judgment and hence being an unsuitable candidate for attribution to the agent herself, unconscious competition for attention looks *prima facie* relevant both to an agent's evaluative stance and her responsiveness to reasons. Think of how natural it would be to say of the wife that she

[15] See Carruthers 2015 for an extended argument that unconscious *decisions* to redirect attention genuinely deserve to be described as such. They are events of the same type, realized in the same brain networks, as the decisions that issue from conscious forms of reasoning and deliberation.

doesn't care enough about her anniversary, and for her spouse to blame her for forgetting. Indeed, if she cared enough or thought it important enough, it would be more likely that the cues she encounters would be judged significant enough to attract attention and enter consciousness—which is just to say that if she cared enough, she wouldn't have forgotten.

Of course, accepting a role for unconscious judgments of the sort we have outlined isn't sufficient to settle questions of responsibility. That she doesn't attend to the activated belief that today is her anniversary doesn't by itself show that she fails to care enough about her marriage or her spouse. All it shows is that, under those various circumstances in which the belief was activated, it failed to direct her attention when competing against her current goals.

Nonetheless, the nature of such competition should plausibly inform our judgments about (and degree of) her blameworthiness for forgetting. Indeed, it will be consistent with our ordinary assessments of such cases. If she has been harried at work and preoccupied with important matters, then we might *expect* her attention to be primarily directed at those representations, rather than being easily shifted to her anniversary, especially if the latter is particularly non-salient this year (say, it is their 12th anniversary). It is precisely such considerations that could serve to mitigate her blameworthiness for forgetting, or would at least be taken to support understanding on the part of an aggrieved spouse. In contrast, we wouldn't as easily expect her attention to continue to be held by more trivial matters, like her fantasy football league rosters. And, likewise, such considerations wouldn't plausibly excuse at all.

Of course, forgetting an anniversary, while significant, is not especially morally serious (although this might vary with particular relationships). Tragically, people do forget about more important things, such as leaving a child unattended in a hot car. It is implausible to suppose such parents fail to care about their children. After all, they are sincerely grief-stricken when realizing what they have done. The unsettling implication of the role of unconscious attitudes in the competition for attention, however, is that it may well be the case that such parents did fail to have proper concern for their children *under the circumstances*. If the belief that the child was still in the car was activated (for example, when they used their keys to open their office door) but attention was nevertheless firmly focused on the meeting they were running late for, then the fact that there was risk to the child was not assessed as significant enough to redirect their attentional focus.[16]

[16] Much will depend on the particular facts in cases of this sort. It is consistent with forgetting one's child in the car that no belief about the *risk of harm* to the child is activated (even if the presence of the child in the car is). If not activated, then we agree that failing to direct attention to the belief that the child is still in the car doesn't reflect any agential judgment about relative

Implicit bias

The last twenty years has seen an explosion of interest in forms of implicit (unconsciously-operating) bias towards members of social groups (Banaji & Greenwald 2013). There are two main forms of implicit bias. One is affective in nature, such as negative feelings (e.g. disliking) towards black people. The other is cognitive, and includes a variety of stereotypes, such as "men are leaders", "women are caring", "Asians are good at math", and so on. Measures of implicit attitudes are known to dissociate from people's overt attitudes, and apparently predict some real-world behavior independently of the latter.[17]

The question for us is whether implicit attitudes provide a locus of responsibility. The answer may depend, in part, on the best account of the nature of those attitudes. On one view, they are low-level associations among properties, such as between *black* and *bad* or between *woman* and *caring* (Gawronski & Bodenhausen 2006). As such, they wouldn't be characterized by Levy (2014) as personal-level, and hence he would say that one cannot be held responsible for their effects. On another view, in contrast, implicit attitudes are just attitudes like any other (Carruthers 2018). The difference is merely that other influences mostly swamp the effects of implicit biases on one's verbal behavior (e.g. when asked how one feels about black people, or whether one would endorse the claim that women are caring). Other influences would include one's egalitarian beliefs, one's desire to avoid social criticism, a desire to protect one's positive self-image, and so on. One consideration in support of this view is that the gap between implicit and explicit measures of one's attitudes is significantly reduced when one is required to answer swiftly, or when one has to answer while under cognitive load of some sort (Hofmann et al. 2005). Under such conditions one's implicit attitudes are more likely to break through and find expression in one's speech.

On the account of implicit stereotypes developed by Carruthers (2018), for example, stereotypes are just generic beliefs, like "birds fly" or "mosquitoes bite". Generics seem to be the mind's default mode

---

importance. Of course, in most actual cases we won't have sufficient evidence about which representations were activated to draw firm conclusions.

[17] For a meta-analysis suggesting that the real-world effects of implicit attitudes are only minor ones, see Oswald et al. 2013. For a careful and measured reply, see Greenwald et al. 2015, who point out that even minor effects can multiply over time, as well as across large groups of individuals. They also point out that many of the null results included in Oswald et al.'s meta-analysis concern effects that should not have been theoretically predicted, such as testing for approach or avoidance behavior (which are among the predicted effects of *affective* attitudes) following measures of implicit stereotypes (which are *cognitive* in nature).

of generalizing. They are acquired extremely swiftly (especially for negative properties), and even "some", "most", and "all" statements are apt to be stored and recalled as generics (Leslie & Gelman 2012; Leslie 2015). While most people will assent to "mosquitoes bite", even though they know that strictly speaking only female mosquitos do, they are much less likely to assent to "black men are dangerous" or "women aren't leaders". But this isn't because of any intrinsic difference in the beliefs in question. It is rather that one's answers, in the latter cases, are moderated by other attitudes. No one thinks we need to be fair to mosquitoes or to give male mosquitoes their due. But many of us think that individual people should be treated as such, and that one shouldn't form expectations about them in advance. Nevertheless, if one has acquired the relevant stereotypes, then one *will* form such expectations when one is responding unreflectively.

If implicit attitudes are just attitudes that one doesn't consciously acknowledge, then it is obvious that actions that manifest such attitudes will reveal something about one's beliefs and values; and as such, those actions may be ones for which one can be held responsible. But what if implicit attitudes are merely associations among ideas? Here, matters may be more complicated. If exposure to a black face causes a negative affective response in a subject, this implies that she negatively evaluates that black face. To the extent that affect is a product of one's evaluative systems, therefore, even low-level associations might reflect an agent's (perhaps unmediated) evaluations. Such affects may be relevantly indistinguishable, however, from similar affective responses to, say, foods the subject dislikes. Of course, there is arguably an important moral difference between disliking black people, on the one hand, and disliking broccoli, on the other. Such feelings may nonetheless still tell us something about the agent's values.

Regardless of whether we can be responsible for our implicit attitudes (understood as associations) directly, however, holding individuals responsible for the actions that result seems much less problematic. After all, complicated actions, like hiring decisions, are not plausibly driven solely by some low-level association between ideas. Indeed, an extensive range of other representations and contents will be activated, many of which will be globally broadcast, throughout any responsible deliberative process. A hiring agent who rejects an otherwise qualified candidate out of an associated dislike or for a perceived lack of leadership qualities (when these associations fail to fit the facts) misevaluates the candidate. Such errors look to be reasonable candidates for responsible (indeed, blameworthy) action. Knowing about these biases, for example, would provide one with good grounds for critiquing the agent's evaluations of the candidates,

and perhaps for sanctioning her accordingly.[18]

6. What is left for consciousness?

While we have critiqued the consciousness condition on responsibility, we concede that consciousness is often *important* for responsibility. Decisions that are taken following conscious reflection will implicate a wider range of one's beliefs and values than decisions that are spontaneous and that manifest the workings of unconscious processes of one sort or another. This is because spontaneous, intuitive, decisions will generally result from a narrow range of values and information. For example, a swiftly-taken decision not to hire a black candidate because "he doesn't feel right for the job" may result simply from an underlying affective bias against blacks. Had the person reflected further, however, a wider range of attitudes would have come into play, including egalitarian beliefs, strongly felt affective reactions against injustice, and so on. Indeed, there is surely a world of difference between a harried and rushed interviewer who takes an unreflective spur-of-the-moment (but biased) decision, and someone who reflectively decides not to hire the black candidate because he doesn't like blacks and doesn't want any to work in his company. That there is a difference, however, is consistent with claiming that both are blameworthy.

It seems to us obvious that reflective deliberation often increases one's blameworthiness for an action of the very same type as one performed spontaneously and unreflectively. If I carefully and consciously calculate the precise words with which to injure you most I am more blameworthy than if I unthinkingly snap at you after a hard day's work. And this can be true even if what is said is exactly the same. Though we have argued that consciousness is not necessary for responsibility, this does not entail that it is irrelevant. On the contrary, consciousness plays an important and morally significant role in action production, though its role is nonetheless indirect.

It is by no means obvious that consciousness *always* increases one's degree of responsibility for an action, however. One sort of exception might be cases where one's careful and calculated conscious decision serves to hide one's true feelings, which might be elicited in spontaneous, non-deliberative action. Another might include cases of willful ignorance, where one purposely blinds oneself to potentially relevant considerations. In the former, one might be more blameworthy for the unconscious action, and in the latter there may be no difference. (Whether or not the latter involves "tracing," which we explicitly left to the

[18] What count as appropriate sanctions here will depend on further moral (and non-moral) considerations apart from just her blameworthiness.

side, is contentious; see King 2017; Sarch 2016.)

7. Moving forward

If unconscious evaluation and decision-making are prevalent and relevant in the ways we have indicated here, it suggests possible implications for other topics. Here we pursue just one: the traditional understanding of *mens rea* in criminal liability.

To be criminally liable requires performing prohibited conduct (*actus reus*) with a culpable mental state (*mens rea*). It isn't enough to kill someone, say. One must do so purposely, knowingly, recklessly, or negligently.[19] Traditionally, recklessness is understood as the minimum standard of culpability for serious offenses, and conscious awareness marks the dividing line between it and negligence.[20] In cases of both recklessness and negligence, the defendant's conduct poses a substantial and unjustifiable risk. What distinguishes recklessness from negligence is that the reckless agent is aware of that risk, but disregards it, whereas the negligent agent is not aware of the risk (but should have been).

The emphasis on conscious awareness is meant to support and justify the grades of *mens rea*. The involvement of consciousness is required for serious offenses because it tracks the degree of culpability. A defendant who consciously *tries* to kill the victim is more culpable than one who is only conscious of the fact that something he is doing for other reasons *will* kill the victim, who is more culpable than one who is only conscious of ignoring a substantial risk that he will kill the victim. And all of these defendants are more culpable than one who isn't aware that he risks killing the victim, though he should have been so aware. Thus, conscious awareness is thought to track culpability, which in turn helps underwrite the use of graded *mens rea* categories for distinguishing between more and less serious offenses.

If unconscious agential evaluation is possible, however, then the use of consciousness to mark this divide appears misguided. Consider a very basic legal requirement to respect the legitimate interests of others. I am obligated to exercise a degree of care in my actions, so as not to risk serious harm to those around me. Recklessness, then, occurs when I consider how my actions might affect the interests of others in ways that ought to constrain my action, but I proceed with the action anyway. We submit that it is an open question whether such disregarding must be conscious. For it may be that the knowledge that one's

[19] Here we follow the Model Penal Code's gradations of culpability (MPC, sec. 2.02), which sought to improve on the common law's mental state categories. See American Legal Institute 1985.

[20] The traditional understanding in common law also indicates that conscious awareness marks a divide in terms of 'malice'.

action will risk harm to others is activated, but then deemed not relevant enough to one's current concerns to attract attention and become conscious. This case strikes us as falling somewhere between the traditional categories of recklessness and negligence. It may not be as culpable as a case where someone is consciously aware of the risk to others, but dismisses it. It is surely more culpable, however, than a case where knowledge of the risk to others is never activated at all. Unsurprisingly, then, consciousness may be important though not necessary for serious criminal liability (for which recklessness generally serves as the minimum standard), just as it is for moral responsibility.

References

Alloway, T. & Alloway, R., editors (2013). *Working Memory: The connected intelligence*. New York: Psychology Press.

American Legal Institute. (1985). *Model Penal Code*: *Official Draft and Revised Comments*.

Awh, E., Belopolsky, A., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16, 437-443.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4, 829-839.

Banaji, M. & Greenwald, A. (2013). *Blindspot: Hidden biases of good people*. New York: Delcorte Press.

Bassetti, C., Vella, S., Donati, F., Wielepp, P., & Weder, B. (2000). SPECT during sleepwalking. *Lancet*, 356, 484-485.

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

Broughton, R., Billings, R., Cartwright, R., Doucette, D., Edmeads, J., Edwardh, M., Ervin, F., Orchard, B., Hill, R., & Turrell, G. (1994). Homicidal somnambulism: A case report. *Sleep*, 17, 253-264.

Carruthers, P. (2015). *The Centered Mind: What the science of working memory shows us about the nature of human thought*. Oxford University Press.

Carruthers, P. (2018). Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology*, 9, 51-72.

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

Corbetta, M. & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3, 201-215.

Corbetta, M., Patel, G., & Shulman, G. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron*, 58, 306-324.

Darley, J. & Batson, D. (1973). 'From Jerusalem to Jericho': A Study of Situational and Dispositional Variables in Helping Behavior. *Journal of Personality and Social Psychology*, 27, 100-108.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering how the brain codes our thoughts*. New York: Viking Press.

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

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

Fox, K., Spring, R.N., Ellamil, M., Andrews-Hanna, J., & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage*, 111, 611-621.

Gawronski, B. & Bodenhausen, G. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132, 692-731.

Gazzaley, A. (2011). Influence of early attentional modulation on working memory. *Neuropsychologia*, 49, 1410-1424.

Goodale, M. (2014). How (and why) the visual control of action differs from visual perception. *Proceedings of the Royal Society B*, 281, 20140337.

Herdova, M. (2016). What you don't know can hurt you: situationism, conscious awareness, and control. *Journal of Cognition and Neuroethics*, 4, 45-71.

Hodgson, D. (2012). *Rationality + Consciousness = Free Will*. Oxford: Oxford University Press.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31, 1369-1385.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy*, 43, 274-306.

Januszko, P., Niemcewicz, S., Gajda, T., Wolynczyk-Gmaj, D., Justyna, P., Gmaj, B., Piotrowski, T., & Szelenberger, W. (2016). Sleepwalking episodes are preceded by arousal-related activation in the cingulate motor area: EEG current density imaging. *Clinical Neurophysiology*, 127, 530-536.

Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

King, M. (2009). The problem with negligence. *Social Theory and Practice*, 35, 577-595.

King, M. (2014). Traction without tracing: a (partial) solution for control-based accounts of moral

responsibility. *European Journal of Philosophy*, 22, 463-482.

King, M. (2017). Tracing the epistemic condition. In Robichaud, P. & Wieland, J.W. (Eds.), *Responsibility: The Epistemic Condition*. Oxford University Press.

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

King, M. & May, J. (2018). Moral Responsibility and Mental Illness: A Call for Nuance. *Neuroethics*, 11, 11-22.

Kurzban, R., Duckworth, A., Kable, J., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, 36, 661-679.

Leslie, S-J. (2015). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, 114, 393-421.

Leslie, S.J. & Gelman, S. (2012). Quantified statements are recalled as generics. *Cognitive Psychology*, 64, 186-214.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, 6, 47-57

Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Mavda, A. (2018). Implicit bias, moods, and moral responsibility. *Pacific Philosophical Quarterly*, 99, 53-78.

Mele, A. (1995). *Autonomous Agents*. New York: Oxford University Press.

Menon, V. & Uddin, L. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, 214, 655-667.

Milner, A. & Goodale, M. (1995). *The Visual Brain in Action*. Oxford University Press.

Miyake, A. & Shah, P., editors (1999). *Models of Working Memory*. Cambridge University Press.

Pereboom, D. (2016). Omissions and different senses of responsibility. In Buckareff, A., Moya, C., & Rosell, S. (Eds.). *Agency and Moral Responsibility*. Palgrave-Macmillan, 179-191.

Plazzi, G., Vetrugno, R., Provini, F., & Montagna, P. (2005). Sleepwalking and other ambulatory behaviors during sleep. *Neurological Science*, 26, 193-198.

R v Parks. (1992). 2 S.C.R. 871.

Sarch, A. (2016). Equal culpability and the scope of the willful ignorance doctrine. *Legal Theory*, 22, 276-311.

Schlosser, M. (2013). Conscious will, reason-responsiveness, and moral responsibility. *Journal of Ethics*, 17, 205-232.

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

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

Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford University Press.

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

Smith, A. (2012). Attributability, answerability, and accountability: in defense of a unified account. *Ethics*, 122, 575-589.

Smith, H. (1983). Culpable ignorance. *Philosophical Review* 92, 543-571.

Smith, H. (2011). Non-tracing cases of culpable ignorance. *Criminal Law and Philosophy*, 5, 115-146.

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

Strawson, G. (1986). *Freedom and Belief*. Oxford University Press.

Tamber-Rosenau, B., Esterman, M., Chiu, Y-C., & Yantis, S. (2011). Cortical mechanisms of cognitive control for shifting attention in vision and working memory. *Journal of Cognitive Neuroscience*, 23, 2905-2919.

Terzaghi, M., Sartori, I., Tassi, L., Rustoni, V., Proserpio, P., Lorusso, G., Manni, R., & Nobili, L. (2012). Dissociated local arousal states underlying essential clinical features of non-rapid eye movement arousal parasomnia: an intracerebral stero-electroencephalographic study. *Journal of Sleep Research*, 21, 502-506.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. (2015). Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience*, 27, 853-865.

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

Wieland, J.W. & Robichaud, P. (2017). Blame transfer. In Robichaud, P. & Wieland, J.W. (Eds.), *Responsibility: The Epistemic Condition*. Oxford University Press, 281-298.

Williams, L.E. & Bargh, J. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 24, 606-607.