Philip Robichaud

Some Doubts About the Consciousness Requirement for Moral Responsibility

Neil Levy has a habit of churning things up for philosophers and others interested in issues of moral responsibility. His probing and careful work on the pervasiveness of responsibility-undermining luck in *Hard Luck* revealed that there remains much rich and unexplored terrain in a field that is often thought to be fully-charted. And just as philosophers were getting acquainted with and developing responses to this work, we were met with another book-length exploration of moral responsibility that is no less provoking. Although the positions defended in *Consciousness and Moral Responsibility* are aimed at similar aspects of the debate, there is no redundancy here. We are again met with novel and compelling arguments that constitute a refreshing reset of many prominently held positions in the field. Moreover, this work is a real exemplar of the kind of scholarship that takes lessons from the social psychological and cognitive sciences and brings them to bear on philosophical questions. I agree with much of what Levy argues in this book. In particular, I share his position that agents who lack creature consciousness and many who act on implicit racial, sexist, or other biases fail to meet one or another condition of moral responsibility. However, I was able to locate a few of Levy's moves that I found less convincing than others. In Section 1, I take a close look at his formulation of the consciousness thesis, which states that in order for an agent to be morally responsible for her action she must be conscious of those facts in virtue of which her action is morally significant. I argue that Levy's account of what precisely

Correspondence:
Philip Robichaud, Delft University of Technology, Netherlands.
Email: p.j.robichaud@tudelft.nl

Journal of Consciousness Studies, **22**, No. 7–8, 2015, pp. ?–?

agents must be conscious of in order to satisfy the consciousness thesis has the problematic feature that it cannot account for blameworthiness attributions in cases where agents act from false belief that their action is wrong. In Section 2, I examine Levy's arguments against real self theories of moral responsibility. Here I develop an objection to his central premise that attitudes about the moral significance of an agent's action must be conscious if they are to interact sufficiently with the relevant elements of the agent's real self. I also argue that the kinds of non-conscious attitudes that inform real self theories lack the defective thinness of content that Levy ascribes to them. In Section 3, I explore Levy's argument that control-based theories of moral responsibility that require agents to be regularly receptive to reasons are committed to the consciousness thesis. I argue that this argument is unsound on the most plausible account of reasons-receptivity according to which deliberation can be rational and reasons-receptive even when it occurs beneath the surface of awareness.

1. Clarifying the Consciousness Thesis

Levy's task in Chapter 2 is to provide a precise definition of the consciousness thesis (CT). After considering and then discarding several options (including, A is conscious of some attitude P iff A is occurrently tokening P; iff A is dispositionally aware of P; iff P is online), Levy comes to the view that the kind of consciousness that is relevant to CT is personal availability (Levy, 2014, pp. 31–4). More precisely, it's the view that A is conscious of some attitude P iff A has the capacity effortlessly and easily to retrieve P and P is online. A word about each conjunct. According to Levy, an attitude is easily and effortlessly retrievable if a sufficiently large set of general cues would occurrently token it, and it is online when it is actually guiding behaviour. Thus, Angela Smith's forgetful friend (2005, p. 236) isn't conscious of the fact that it is her friend's birthday because her belief is neither effortlessly nor easily retrievable. General cues, such as seeing a phone or a calendar, fail to occurrently token it. Having argued for a personal availability account of consciousness, Levy turns to the question of what information or content agents must be aware of, and this is where I would like to focus some critical attention.

Levy argues that we must be conscious of information about facts that explain the valence of responsibility (Levy, 2014, p. 36). For

example, the fact that Zoë's action causes easily avoidable harm explains why the valence of her responsibility is negative (i.e. Zoë would be blameworthy). Levy rightly notes, however, that the valence of responsibility attributions tracks information about what the agent takes to pertain rather than what in fact pertains. Thus, it is the fact that Zoë takes the action to be the cause of avoidable harm that explains why she is blameworthy. This subjectivization or internalization of the responsibility-relevant information allows Levy to account for cases in which an agent takes herself to be performing an action that is wrong or bad, even though they are actually bringing about some good, as would be the case if Zoë donated to an Oxfam collection box thinking it was 'People for Cannibalism' (*ibid*., p. 37). She is blameworthy but not because her action was bad. Rather, her blameworthiness rests on the fact that she took herself to be performing an act that would have had bad-making features if her beliefs about it were true. One final and crucial feature of Levy's account is that agents must be conscious of information that explains the valence of responsibility 'under an appropriate description', where the appropriate description is one that 'captures some aspect of the badness of the action' (*ibid*., p. 37). Levy must include this in order to rule out the possibility that CT would require an agent to be conscious of her action under every relevant description. Agents need not be aware of any description of their action that is either additional or redundant.

Now, there seems to be a tension between what Levy says agents must be conscious of (namely, facts that explain the valence of responsibility attributions) and his account of when descriptions of these facts are appropriate. Again, someone who Φ's is blameworthy only if she (1) takes herself to be performing Φ under description D and (2) D captures the badness of Φ. That is, she is blameworthy only when the aspect of the action of which she is conscious is identical with the aspect that actually is a bad-making feature of the action. This much is fine, on the assumption that the relevant bad-making features are *not* determined by what the agent takes to pertain. Let's assume that Jan believes truly that he is donating to People for Cannibalism and that his action has its bad-making features independently of what Jan believes. So, if we want to know whether Jan's donation is blameworthy, we need to know (1) that he takes himself to be performing an act with the description 'I'm donating to People for Cannibalism' and (2) this description captures the non-subjective badness of such a donation. Since the bad-making features of Jan's action involve his giving support to immoral, non-consensual

cannibalistic practices, we have the requisite match between the content of his conscious attitude and the bad-making features of the action. The description under which he acts is appropriate.

Things are less clear-cut in cases where the valence of responsibility turns on what the agent believes. Levy maintains that an agent can be blameworthy in cases where she falsely believes that her action has features that would make it bad, even when the act is not bad in itself. So, imagine again that Zoë takes herself to be donating to People for Cannibalism, but she's actually donating to Oxfam. In order to figure out whether she is conscious of the morally significant feature of her action under the appropriate description, the description mustn't capture a feature of her action — giving to Oxfam is great, after all. Instead, the fact that explains Zoë's blameworthiness and that constitutes the moral character of her action is that she takes herself to be donating to People for Cannibalism. It follows that Zoë is conscious of her action under the appropriate description only if the description captures the fact <that she believes herself to be donating to People for Cannibalism>. But, given that this is a fact about her and not her action, it simply cannot be the case that Zoë's beliefs about what she is doing under some description will capture the fact that makes her action blameworthy. In order to do so, she would need to believe that she is doing something that she believes to be donating to People for Cannibalism, and there is little reason to think she would ever be conscious of such a baroque belief at the time of her action. It follows that for any case similar to Zoë's, wherein an agent acts from the false belief that she is performing an action that, if her beliefs were true, would have bad- or wrong-making features, agents would almost certainly not satisfy the CT — they would only if, for some odd reason, they hold the appropriate baroque belief. Since there is significant intuitive pressure in the direction of thinking such agents can be blameworthy, Levy's criterion for the appropriateness of the action description should be modified.

2. Implicit Attitudes and Real Self Views

Levy's argument against the first class of real self theories of moral responsibility is quite simple (*ibid*., pp. 92–5). These theorists hold:

> (RS) A is responsible for action Φ only if Φ expresses A's evaluative agency.

Levy will defend:

> (EAC) Φ expresses A's evaluative agency only if A is conscious of the evaluative significance of Φ.

It follows form RS and EAC that:

> (CT) A is responsible for action Φ only if A is conscious of the evaluative significance of Φ.

In defence of EAC, Levy makes the following argument:

(1) Φ expresses A's evaluative agency only if the attitudes that cause Φ (and give it its moral significance) interact with A's personal-level concerns or goals.
(2) The attitudes that cause Φ interact with the A's personal-level concerns or goals only if A is conscious of the evaluative significance of Φ.
So, EAC.

Premise (1) of this sub-argument states that there must be some relationship between the attitudes that cause the action and the agent's agential self, which is here understood as being constituted in part by personal-level concerns or goals. Levy's proposal is that attitudes that express evaluative agency must have been checked against the agent's personal-level concerns. Personal-level concerns are what constitute an agent's evaluative stance, and since stances must be coherent and reasonably consistent, an action caused by attitudes that either fail to cohere or aren't consistent with personal-level attitudes cannot be rightly said to express the agent's evaluative stance.

The rationale for (2) is that there seems to be no imaginable way of assessing non-conscious attitudes for their coherence or consistency with personal-level concerns. Unless, of course, the evaluative significance of A was both easily and effortlessly retrievable and online. Only then does it seem plausible that an attitude could interact sufficiently with the agent's personal-level concerns.

I will focus my critical attention in this section on premise (2). It seems that the claim that agents can assess an attitude for coherence or consistency with personal-level concerns or goals only when the attitude is conscious is in some tension with the position Levy takes toward the question of whether rational deliberation requires consciousness of cognitive machinations that constitute it (*ibid*., p. 23). Regarding the latter, he argues that it is not a necessary condition of rational deliberation that an agent be aware of the processes involved in (1) the discovery of her reasons' weights, (2) the weighing of

reasons against each other, and (3) the coming to a decision on the basis of this weighing. One of Levy's reasons for thinking this, aside from the fact that it seems phenomenologically accurate, is that deliberation so understood is a process of *discovery* of certain facts about our reasons. We need not be consciously deliberating in order for these processes to occur. But, if weighing reasons and coming to decisions on the basis of such weighings is a process that can take place beneath the surface of awareness, then it is natural to wonder why, as (2) implies, an agent wouldn't be able to assess the consistency and coherence of certain non-conscious attitudes with the set or relevant subset of attitudes that make up her evaluative stance.

Consider Oliver Single, a John La Carré character to whom Nomy Arpaly appeals as an example of an agent who is morally praiseworthy despite acting from non-conscious attitudes (Arpaly, 2002, p. 4). It seems no less plausible to think that the non-conscious attitude that caused Single to pick up the phone survived many instances of non-conscious interaction with other of his personal-level attitudes, than it is to think that a different agent, call him Double, was able to non-consciously both weigh reasons in favour and against and come to the same decision. In the latter process, Double must discover the relative strengths of his practical reasons and decide to reach for the phone on the basis of his strength assessment. In the former process, Single's attitude is tested either sequentially or serially against other beliefs and consistency is established (or not). The cognitive operations involved in both kinds of process are complex and extended in time and they both seem like acts of discovery. In the one case it is discovery of the strength of practical reasons, and in the other it is discovery of the consistency and coherence of attitudes. They also both involve operations on contentful cognitive states that are in part determined by the content of these states. Given that there are many similarities between these two processes, one wonders why Levy would maintain that we can successfully engage in the one non-consciously, but not the other.

It is of course consistent with maintaining this line of questioning that consciousness of the moral significance of attitudes is required for *certain* actions. For example, certain attitudes may be related to a sufficiently large number of personal-level concerns that the calculative demands on assessing consistency would require that the attitude be broadcast globally. It is also the case, of course, that for some similarly complex deliberations, consciousness of the deliberative goings-on may promote rational decision making. The burden on

Levy's interlocutor, though, is only to show that there are some, perhaps limited, cases in which attitudes can be assessed for consistency with personal-level attitudes via non-conscious processes. The similarities between this assessment procedure and run-of-the-mill non-conscious deliberation suggest that this burden can be met.

Levy's next targets are Scanlon's and Arpaly's more moderate real self views (Levy, 2014, pp. 96–102). Scanlon (2002) argues both that non-conscious attitudes can be reasons for action and that they are sensitive to judgments, and Arpaly (2002) argues that actions caused by non-conscious attitudes can express the agent's level of moral concern, where this falls short of expressing evaluative agency. Levy argues that the problems with this class of real self views stem largely from the content of non-conscious attitudes. Of import is the fact that implicit attitudes, such as those discussed by Uhlmann and Cohen (2007; 2005), are formed when the brain associates a certain stimulus with a reward or punishment, which then leads to the formation of dispositions to have valenced emotional responses to the presence of the stimulus (Levy, 2014, p. 98). When the emotional response is positive, agents want the reward that was associated with the stimulus. Levy points out that the desire for reward may persist even in the absence of the dispositions that normally constitute something as reason-giving. So, *contra* Scanlon, implicit attitudes may incline an agent toward a certain action without it being reason-giving at all. Levy also argues that implicit attitudes are notoriously judgment *in*sensitive (*ibid*., p. 99). As such, implicit attitudes can only be altered indirectly, they don't function as justificatory reasons for the agent, and, indeed, they actually aren't beliefs at all. Finally, Levy argues that the fact that moral concern or responsiveness to moral reasons plays no role in the acquisition or persistence of implicit attitudes, they fail on Arpaly's account to be appropriate grounds for judgments of blameworthiness (*ibid*., p. 100). The lesson Levy draws is that Scanlonian and Arpalian real self accounts cannot establish moral responsibility for actions caused by implicit attitudes.

These arguments are persuasive, and I will not dispute them directly. What I will argue, however, is that the problematic features of implicit attitudes are not shared by other important classes of non-conscious attitudes that have been the concern of real self theorists. Consider, again, the actions of Single. The non-conscious attitudes that lead him to defect are not the effects of associations that manifest in a disposition to defect when opportunity presents itself. Rather, on Arpaly's telling at least, Single's attitudes are the result of his dis-

affection over time with the activities of his firm. Though he never reasons or deliberates consciously about what he should do in response to this disaffection, his unconscious reasons for acting are not plausibly thought to be the effect of anything like the prediction error systems that Levy suggests are crucial to the formation of implicit attitudes. In part because of their different aetiology, Single's attitudes may not share in the defective thinness of content that plagues implicit associations between, say, being male and being a police chief. In fact, Single seems to have the content-ful though non-conscious belief that he should defect, a belief that seems to be both sensitive to judgments and the product of moral concern — his lingering disaffection can only be explained by some degree of moral reasons-responsiveness. A similar analysis can be offered for the case of Huck. His non-consciously tokened respect for Jim's humanity is what drives him. This attitude also has morally thick content, and we are to imagine that its aetiology involves Huck's repeated displays of moral concern. These reflections suggest that, although Levy has presented a powerful case against the position that agents can be directly morally responsible for actions that are grounded in implicit attitudes of the kind that lead daily to countless instances of racist and sexist behaviour, there is still reason to think that the features of such actions that recommend this unpalatable conclusion are not shared by certain central cases of agency driven by non-conscious attitudes.

3. Control Theorists, Deliberation, and Reasons-Receptivity

The central question discussed in Chapter 6 is whether control over one's actions requires awareness of their moral significance. Levy focuses his enquiry by taking up Fischer and Ravizza's (2000) influential account of guidance control and its appeal to reasons-responsive mechanisms. Moderately reasons-responsive mechanisms must be regularly receptive to reasons, which means that the mechanism must produce an understandable (to a third party) pattern of reasons-responsiveness. Levy argues that agents such as Ken Parks who are suffering from global automatism fail to be regularly receptive given that the scripts from which they act fail to respond to significant swaths of reasons (Levy, 2014, pp. 111–3). Agents who lack creature consciousness are only capable of stereotypy, which falls well short of controlled action necessary for moral responsibility. Levy moves on to consider creature conscious agents who are unaware of the moral

significance of their actions. Here the argument is that the kind of flexibility required for regular, patterned receptivity to reasons can only be present if the agents are conscious of both the attitudes that cause a given action as well as its moral significance (*ibid*., pp. 115–6).

Levy's argument in this section is:

(1) A is morally responsible for action Φ only if A has guidance control over Φ.
(2) A has guidance control over Φ only if A is regularly receptive to reasons, including any moral reasons relevant to Φ-ing.
(3) A is regularly receptive to reasons, including any moral reasons relevant to Φ-ing only if A is aware both of the facts that lead her to Φ and the moral significance of her Φ-ing.

So, A is morally responsible for action Φ only if A is aware both of the facts that lead her to Φ and of the moral significance of her Φ-ing.

Premises (1) and (2) are taken on board for the purposes of argument. The rationale for (3) is that, absent awareness of the facts that shape our actions, we are also unaware of the moral significance of our actions and, thereby, unaware of the moral reasons relevant to our actions. And, lacking such awareness seems straightforwardly to be incompatible with being receptive to the relevant moral reasons.

I submit that Levy's defence of (3) relies on an idiosyncratic interpretation of reasons-receptivity. Fischer and Ravizza argue, I think convincingly, that receptivity to reasons is at bottom a matter of forming beliefs that can function as reasons for action (Fischer and Ravizza, 2000, p. 90, n. 35). In order to determine whether a mechanism is receptive to sufficient reasons to Φ, one needs to check whether the agent would form the belief that she should Φ in worlds where she has sufficient reasons to Φ. Importantly, it seems not to be a requirement that a reasons-receptive agent hold this belief occurrently at any time. To see how this is possible, return again to the reflections above about episodes of non-conscious deliberation. When we deliberate, we employ mechanisms that are (we hope) sufficiently responsive to reasons. In Levy's terms, these mechanisms allow us to discover our practical reasons and assess their comparative weights. When these mechanisms recognize reasons to be sufficient, we, if rational, form the intention to act accordingly. Moreover, these episodes of reasoning seem to exhibit understandable patterns of receptivity. If they didn't, and if the only way of securing regular receptivity was via

conscious awareness of the moral significance of the available options and their comparative weights, then consciousness of certain facts would be a requirement on rational deliberation. This is an implausible view that both Levy and his interlocutors are right not to adopt. If it is relatively uncontroversial that rational deliberation as I've just construed it can be shielded from our conscious awareness, and if in deliberating we can employ regularly receptive, reasons-responsive mechanisms, then premise (3) will turn out to be false. Of course, it doesn't follow from this that agents who act on implicit biases are indeed capable of acting on moderate reasons-responsive mechanisms. Subjects in Uhlmann and Cohen's study may fail to act on a regularly receptive mechanism for precisely the reasons that Levy adduces (having to do with acting on the basis of scripts that are only narrowly and inflexibly reasons-responsive). However, it does not follow from the fact, if it is a fact, that agents who act on implicit biases are not acting on moderately reasons-responsive mechanisms that all cases of non-conscious deliberative action lack responsibility-relevant reasons-responsiveness.

4. Conclusion

In this commentary I called for a revision to Levy's formulation of the consciousness thesis, and I offered some reasons to think that Levy's primary targets (namely, real self and control theorists of moral responsibility) can resist it. If my arguments are successful, then agents who act from certain non-conscious attitudes may be morally responsible after all. Although my attention in this commentary has been mainly critical, I believe that Levy's latest book is commendable for many reasons. To enumerate them would take me well beyond the scope of this paper. Most prominently, in my view, Levy offers the reader an opportunity to think carefully about the way in which our ever-deepening understanding of human psychology and action can inform philosophical reflection about some of the most important and enduring moral questions.

References

Arpaly, N. (2002) *Unprincipled Virtue: An Inquiry into Moral Agency*, Oxford: Oxford University Press.

Fischer, J.M. & Ravizza, M. (2000) *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

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

Scanlon, T.M. (2002) Reasons and passions, in Buss, S. & Overton, L. (eds.) *Contours of Agency: Essays for Harry Frankfurt*, Cambridge, MA: MIT Press.

Smith, A.M. (2005) Responsibility for attitudes: Activity and passivity in mental life, *Ethics*, **115** (2), pp. 236–271.

Uhlmann, E.L. & Cohen, G.L. (2005) Constructed criteria redefining merit to justify discrimination, *Psychological Science*, **16** (6), pp. 474–480.

Uhlmann, E.L. & Cohen, G.L. (2007) I think it, therefore it's true: Effects of self-perceived objectivity on hiring discrimination, *Organizational Behavior and Human Decision Processes*, **104** (2), pp. 207–223.