Gregg D. Caruso

If Consciousness is Necessary for Moral Responsibility, then People are Less Responsible than We Think

In *Consciousness and Moral Responsibility*, Neil Levy argues for and defends the *consciousness thesis*, which maintains that 'consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility' (Levy, 2014, p. 1). He contends that since consciousness plays the role of integrating representations, behaviour driven by non-conscious representations are inflexible and stereotyped, and only when a representation is conscious 'can it interact with the full range of the agent's personal-level propositional attitudes' (*ibid.*, p. vii). This fact entails that consciousness of key features of our actions is a necessary (though not sufficient) condition for moral responsibility since consciousness of the morally significant facts to which we respond is required for these facts to be assessed by and expressive of the agent him/herself. Furthermore, he argues that the two leading accounts of moral responsibility — *real self* (or *evaluative accounts*) and *control-based* accounts — are committed to the truth of the consciousness thesis despite what proponents of these accounts maintain. According to Levy: (a) only actions performed consciously express our evaluative agency, and that expression of moral attitudes requires consciousness of that attitude; and (b) we possess responsibility-level control only over actions that we perform consciously, and that control over their moral significance requires consciousness of that moral significance.

Correspondence:
Email: gcaruso@corning-cc.edu

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

In these comments I will grant Levy his main conclusion — i.e. the truth of the consciousness thesis. While most critics are likely to deny this demanding condition (see, for example, Sripada, 2015), I believe Levy makes a strong case for the integrative function of consciousness and its importance to moral responsibility. In my book *Free Will and Consciousness* (Caruso, 2012), I also argue that consciousness is a necessary condition for free will and desert-based moral responsibility. Though the details of our accounts differ in significant ways, we both agree on the central question, and where our accounts differ I am willing to grant Levy the details of his account for the sake of this paper. Rather than question the consciousness thesis, then, I will take the road less travelled and argue that Levy should, given his own commitments, embrace an *even more* sceptical conclusion than the one he adopts.

1. Global Automatism, Implicit Bias, and Situationism

In the concluding section of *Consciousness and Moral Responsibility*, Levy addresses the concerns of theorists like myself (Caruso, 2012) who worry that the ubiquity and power of non-conscious processes either rules out moral responsibility or severely limits the instances where agents are justifiably blameworthy and praiseworthy for their actions. He maintains that adopting the consciousness thesis need not entail scepticism of desert-based moral responsibility since the consciousness condition can be (and presumably often is) met. His argument draws on an important distinction between cases of global automatism and implicit bias, on the one hand, and cases drawn from the *situationist* literature on the other. Levy maintains that in the former cases (global automatism and implicit bias), agents are excused moral responsibility since they either lack creature consciousness, or they are creature conscious but fail to be conscious *of* some fact or reason which nevertheless plays an important role in shaping their behaviour. In situational cases, however, Levy maintains that agents *are* morally responsible despite the fact that their actions are driven by non-conscious situational factors, since the moral significance of their actions remains consciously available to them and globally broadcast (Levy, 2014, p. 132).

To see if Levy is correct about this, let us examine more closely the three types of cases — which I will define as follows:

(1) **Type-1 Cases:** These are cases of *global automatism*, where agents either lack creature consciousness altogether or are in a

very degraded global state of consciousness. These cases are dramatic, puzzling, and relatively rare. Examples include cases of somnambulism such as the Kenneth Parks case.

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

(3) **Type-3 Cases:** Perhaps even more common still are cases where agents are conscious of facts that shape their behaviour, but conscious neither of *how*, nor even *that*, those facts shape their behaviour. Examples of type-3 cases can be found in the *situationist literature* — for example, an agent may be conscious that they previously held a hot cup of coffee, but not conscious *that* (or *how*) the cup of coffee affected their judgment of others (Williams and Bargh, 2008).

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

Let me begin by examining the cases where Levy and I agree. In type-1 cases, agents either lack creature consciousness altogether or are in a very degraded global state of consciousness. In the Kenneth Parks case, for example, Parks was (presumably) in a state of somnambulism when he drove 23 kilometres to his in-laws' house and proceeded to stab to death his mother-in-law and strangle unconscious his father-in-law. Both Levy and I agree that Parks, being driven by *action scripts* in this situation, is not (directly) morally responsible for his actions since he acts without consciousness of a range of facts, each of which gives to his actions moral significance — i.e. he is 'not conscious *that he is stabbing an innocent person*; he is not conscious

that [his mother-in-law] is begging him to stop, and so on' (Levy, 2014, p. 89). Since Parks is not able to assess the significance of his action in the light of his personal-level attitudes, his behaviour does not express his evaluative agency. He also lacks the kind of flexible, reasons-responsive, online adjustment of behaviour that comes with consciousness. It is for these reasons that Parks fails to be morally responsible.

While it's likely that Levy will receive wide agreement on type-1 cases, type-2 cases are more controversial. Consider, for instance, the study by Uhlmann and Cohen (2005) on *implicit sexism*. In this study, subjects were asked to rate the suitability of two candidates for police chief, one male and one female, where one candidate was presented as 'streetwise' but lacking in formal education while the other one had the opposite profile. Despite the fact that Uhlmann and Cohen varied the sex of the candidates across conditions — so that some subjects got a male streetwise candidate and a female well-educated candidate while other subjects got the reverse — subjects considered the male candidate significantly better qualified in *both* conditions. This indicates that the hiring decision was the result of implicit sexism rather than the qualifications of the candidates. While it's tempting to hold agents morally responsible in such situations, Levy provides compelling reasons for resisting this temptation.

What's lacking in the above scenario is consciousness of the facts that give the agent's decision its moral significance. Rather than being conscious of the sexist attitude, the agent is conscious of a confabulated criterion which itself seems plausible — i.e. the importance of being streetwise or highly educated. It is this confabulated criterion that is globally broadcast and assessed against the agent's beliefs, values, and other attitudes. Since the agent is unaware of the implicit sexism, he is unable to evaluate and assess it against his personal-level attitudes. It would therefore be wrong to attribute the sexism to the agent's *real self* or consider it an expression of his evaluative agency.[1] Furthermore, since the agent is conscious neither of the implicit attitude that caused the confabulation, nor of the moral significance of the decision, he is unable to exercise guidance control (or moderate reasons-responsiveness) over either. Given that real self and control-based accounts represent the best candidates we have for necessary conditions for moral responsibility, and given that these conditions are

[1] This is assuming that the agent does not openly endorse sexism in hiring decisions.

not met in type-2 cases, I concur with Levy that we should excuse moral responsibility in these cases as well.

It's at this point that Levy and I begin to part ways. But before turning to type-3 cases, I would just like to say one last thing about type-2 cases. Levy does not speculate on *how* pervasive type-2 cases may be. I wish he had since there is some empirical evidence that implicit biases may be more common than we think.[2] Clearly it's an empirical question the extent to which we are guided by implicit biases, but it would be helpful to know whether Levy thinks such cases are common and people should *generally* be excused moral responsibility or whether he thinks they are rare. One gets the impression from the first two-thirds of his book that Levy believes people are *significantly* less responsible than we might think — especially when he says that adopting the consciousness thesis 'will lead to fewer people being unjustly held morally responsible' (Levy, 2014, p. x). Yet in the final section, when he addresses my concern that the ubiquity and power of non-conscious processes severely limits moral responsibility, one gets a different impression. I would like to give Levy the opportunity, then, to clarify how pervasive he thinks type-2 cases may be since our views may not differ much if he thinks they are rather common.

That said, I think we have a real disagreement regarding type-3 cases. In his concluding section, Levy writes: 'Caruso points to the voluminous evidence that situational factors — deliberately manipulated by an experimenter or simply encountered in the world — may not merely influence our actions but profoundly transform their character. Given one set of primes, we may act morally; given another, immorally' (*ibid.*, p. 131). He goes on to acknowledge that these experiments 'demonstrate that whether an agent does the right thing or not… may be strongly influenced by nonconscious factors (or factors the influence of which agents are not consciousness)' (*ibid.*, p. 132), but he maintains that this concession does not threaten moral

[2] Research has found that implicit biases are pervasive and robust (Greenwald, McGhee and Schwartz, 1998; Kang *et al.*, 2012; Kang and Lane, 2010; Nosek *et al.*, 2007). Everyone possesses them, even people with avowed commitments to impartiality such as judges (Rachlinski *et al.*, 2009). There is even some evidence that implicit attitudes may be better at predicting and/or influencing behaviour than self-reported explicit attitudes (Bargh and Chartrand, 1999; Beattie, Cohen and McGuire, 2013; Ziegert and Hanges, 2005). For an excellent survey of the implicit bias literature — which discusses such biases as racial bias (as evidenced by the shooter/weapons bias, healthcare biases, and the biases exhibited by defence attorneys, juries, and judges) — see, Staats (2014).

responsibility. In the following section, I will challenge this conclusion and I will do so using Levy's own account of why consciousness is necessary for moral responsibility.

2. The Situationist Challenge

Like Levy, I am a free will sceptic primarily for philosophical reasons, not empirical ones. I maintain that philosophical arguments on their own are sufficient for showing that people are never morally responsible for their actions in the basic desert sense — the sense that would make us *truly deserving* of blame and praise. But, also like Levy, I further maintain that factors having to do with consciousness and empirical developments in the behavioural, cognitive, and neurosciences — especially those related to situationism, automaticity, and the adaptive unconscious — represent a separate and unique problem for moral responsibility. It therefore seems that our only disagreement (to the extent there actually is one) is over type-3 cases and the degree to which adopting the consciousness thesis would excuse moral responsibility.

Consider the following examples taken from the situationist literature — each one representing a case in which agents are conscious of facts that shape their behaviour, but conscious neither of *how*, nor even *that*, those facts shape their behaviour.

> **Example 1:** Experiments carried out by Bargh, Chen and Burrows (1996) found that when trait constructs were non-consciously activated during an unrelated task, what is known as priming, participants were subsequently more likely to act in line with the content of the primed trait construct. In one experiment, participants were primed on the traits of either rudeness or politeness (or neither) using a scramble-sentence test in which they were told to form grammatical sentences out of short lists of words. Participants were exposed to words related to either rudeness (e.g. rude, impolite, obnoxious), politeness (e.g. respect, considerate, polite), or neither. Participants were told after completing the test that they were to go tell the experimenter they were done. When they attempted to do so, however, the experimenter was engaged in a staged conversation. Bargh and his colleagues wanted to see if participants would interrupt. They found that among those primed for 'rudeness' 67% interrupted, among those primed for 'politeness' only 16% interrupted, and for the control group 38% interrupted. In addition, during an

extensive post-experiment debriefing, none of the participants showed any awareness or suspicion of the possible influence of the scramble-sentence test on their interrupting behaviour.

Example 2: In a classic study on the influence of mundane physical objects on situational construal and competitive behavioural choice, Kay *et al*. (2004) asked subjects to participate in a financial game. Kay and his colleagues found that those who sat at a table with a briefcase strategically placed on it played the game far more competitively and selfishly than did participants who sat near a backpack. The mere presence of a briefcase, which is presumably associated with business, is enough to trigger behavioural dispositions associated with business. This occurred, once again, without the participants' awareness of the relevant influence. When probed in post-experiment interviews, none of the participants were aware of any aspect of the physical environment that may have influenced their physical strategies.

Example 3: In a classic study on how extraneous factors influence judicial decisions, researchers found that experienced parole judges in Israel granted freedom 65% of the time to the first prisoner who appeared before them on a given day, but by the end of the morning session the chances of receiving parole dropped to almost zero (Danziger, Levav and Avnaim-Pesso, 2011). Disturbingly, they found that the decisions of the judges had less to do with legal reasoning and facts, and more to do with *ego depletion* and what the judges ate for breakfast. They recorded the two daily food breaks taken by the judges, which resulted in segmenting the deliberations of the day into three distinct 'decision sessions'. They found that the percentage of favourable rulings drops gradually from 65% to nearly zero within each decision session and returns abruptly to 65% after a break. These findings reveal that judicial rulings can be swayed by extraneous variables that should have no bearing on legal decisions.

I chose these three examples because they represent a range of different scenarios and make use of different ways in which agents can be influenced by non-conscious factors. If agents fail to be morally responsible in these cases, it would be easy to see how that conclusion could be generalized to other similar cases.

Let's begin with Example 1. In this situation an agent behaves rudely by interrupting a conversation, but does so because of situational factors the influence of which the agent is not conscious. Should the agent be excused moral responsibility? I contend that if we apply the consciousness thesis and the same considerations we did in type-2 cases, we should answer in the affirmative. First of all, it's *prima facie* plausible to think that an agent in this situation fails to be conscious of the facts that give his action its moral significance. It's reasonable to think that rather than being conscious *that he is acting rudely* (under this or a similar description), the agent is instead conscious of some confabulated reason for his behaviour. Like the implicit sexism case discussed above, this would mean that rather than being conscious of the primed trait construct for rudeness, the agent is conscious of a confabulated reason for his behaviour which itself seems plausible. In turn, it would be this confabulated reason that is globally broadcast and assessed against the agent's beliefs, values, and other attitudes. Since the agent is unaware of the primed trait construct for rudeness, he is unable to evaluate and assess it against his personal-level attitudes. Hence, we should conclude that it is not a reflection of his evaluative agency.

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

The second example is similar to the first, except that the situational factor involved is a mundane physical object (a briefcase) rather than a scramble-sentence test. I would argue the same thing here — i.e. that the agent's selfish and competitive behaviour is neither a reflection of their evaluative agency nor something they exercise guidance control over. According to the consciousness thesis, if an action is morally

bad the agent must be conscious of the aspects that make it bad, and conscious of those aspects under an appropriate description, in order to be blameworthy for the action. Yet, in this example, it's again reasonable to think that the agent remains unaware of the morally significant facts that give their action its moral valence. Rather than being conscious *that they are acting selfishly* (under this or a similar description), the agent is conscious of confabulated reasons for playing the game as they do. It is these confabulated reasons that are globally broadcast and assessed against the agent's beliefs, values, and other attitudes. We can say that, while the agent is conscious of the briefcase placed on the table, they are not conscious *how* or *that* the briefcase is triggering various behaviours and dispositions associated with business (i.e. selfishness and competitiveness). Because they remain unaware of this fact, and because this fact is morally significant, we should conclude that the consciousness thesis is not satisfied and the agent is not morally responsible.

Example 3 is a little different. Technically speaking it is not an example of situationism, but it does satisfy my definition of a type-3 case since the judges are conscious of the facts that shape their behaviour (the time of day, when they last eat, etc.), but conscious neither of *how* or *that* those facts shape their behaviour. While the judges believe they are employing objective legal reasoning and facts to reach their decisions, we can see that their decisions are being influenced by powerful extraneous variables that should have no bearing on legal decisions. Recall that at the end of each 'decision session' the percentage of favourable rulings drops to nearly zero. Why is this the case? One explanation, the one offered by the researchers who conducted the study, has to do with *ego depletion* (see Muraven and Baumeister, 2000; Pocheptsova *et al*., 2009; Vohs *et al*., 2008; Gailliot and Baumeister, 2007; Hagger *et al*., 2010). Recent research suggests that making repeated judgments or decisions depletes individuals' executive function and mental resources, which can, in turn, influence their subsequent decisions. For instance:

> [S]equential choices between consumer goods can lead to an increase in intuitive decisionmaking (Pocheptsova et al. 2009) as well as reduced tolerance for pain in a subsequent task (Vohs et al. 2008). Sequential choices and the apparent mental depletion that they evoke also increases people's tendency to simplify decisions by accepting the status quo. German car buyers, for instance, were more likely to accept the default attribute level offered by a manufacturer later in a sequence of attribute decisions than earlier, particularly when these choices followed decisions between many alternatives that had required more mental

> resources to evaluate (Levav 2010). (Danziger, Levav and Avnaim-Pesso, 2011, p. 6889)

In the case of the Israeli judges, what seems to be going on is that, as they advance through the sequence of cases, their executive function and mental resources get depleted (ego depletion) and they become more likely to accept the default, status quo outcome: deny a prisoner's request for parole.

If this is correct, and there appears to be good reason to think it is, we would once again have a failure of the consciousness thesis. Because the judges were conscious neither of the ego depletion nor its effects on their decisions, they could not detect conflicts between their decisions and their personal-level attitudes. What was globally broadcast, and therefore assessed for consistency and conflict, was a confabulated set of standards for parole produced by the ego depletion. The judges were not aware of the default reasoning they were employing as a result of that ego depletion (e.g. *first do no harm*, *maintain the status quo,* etc.). Because they failed to be conscious of the default reasoning, they were unable to assess it against their personal-level attitudes or exercise responsibility-level control over it.

3. Conclusion

In these comments I have tried to argue that, if consciousness is a necessary condition for moral responsibility, then people are less responsible than we think — and less responsible than Levy thinks. Of course, this conclusion may lead some to simply reject the consciousness thesis — one person's *modus ponens* is another's *modus tollens* — but I think that would be a mistake. Levy provides compelling reasons for accepting the consciousness thesis and he makes a strong case for the integrative function of consciousness and its importance to moral responsibility. Rather than question the thesis, then, I have focused my attention on exploring the extent to which adopting it would excuse moral responsibility. I have argued that a wider range of cases than Levy suggests would fail to satisfy the consciousness thesis and that, because of this, moral responsibility would be more limited than we are lead to believe. To what degree Levy actually disagrees with my sceptical extension of his argument is unclear to me, so I very much welcome his response.

References

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

Bargh, J. & Chartrand, T.L. (1999) The unbearable automaticity of being, *American Psychologist*, **54** (7), pp. 462–479.

Beattie, G., Cohen, D. & McGuire, L. (2013) An exploration of possible unconscious ethnic biases in higher education: The role of implicit attitudes on selection for university posts, *Semiotica*, **197**, pp. 171–201.

Caruso, G.D. (2012) *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*, Lanham, MD: Lexington Books.

Danziger, S., Levav, J. & Avnaim-Pesso, L. (2011) Extraneous factors in judicial decisions, *Proceedings of the National Academy of Sciences*, **108** (17), pp. 6889–6892.

Gailliot, M. & Baumeister, R. (2007) The physiology of willpower: Linking blood glucose to self-control, *Personal Social Psychology Review*, **11**, pp. 303–327.

Greenwald, A.G., McGhee, D.E. & Schwartz, J.L.K. (1998) Measuring individual differences in implicit cognition: The implicit association test, *Journal of Personality and Social Psychology*, **74** (6), pp. 1464–1480.

Hagger, M.S., Wood, C., Stiff, C. & Chatzisarantis, N. (2010) Ego depletion and the strength model of self-control: a meta-analysis, *Psychology Bulletin*, **136**, pp. 495–525.

Kang, J. & Lane, K. (2010) Seeing through colorblindness: Implicit bias and the law, *UCLA Law Review*, **58** (2), pp. 465–520.

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

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

Muraven, M. & Baumeister, R. (2000) Self-regulation and depletion of limited resources: Does self-control resemble a muscle?, *Psychology Bulletin*, **126**, pp. 247–259.

Pocheptsova, A., Amir, O., Dhar, R. & Baumeister, R. (2009) Deciding without resources: Resource depletion and choice in context, *Journal of Marketing Research*, **46**, pp. 344–355.

Rachlinski, J.J., Johnson, S.L., Wistrich, A.J. & Guthrie, C. (2009) Does unconscious racial bias affect trial judges?, *Notre Dame Law Review*, **84** (3), pp. 1195–1246.

Sripada, C. (2015) Acting from the gut: Responsibility without awareness, *Journal of Consciousness Studies*, **22** (7–8).

Staats, C. (2014) State of the science: Implicit bias review 2014, *Kirwan Institute*, [Online], http://kirwaninstitute.osu.edu/wp-content/uploads/2014/03/2014-implicit-bias.pdf [2 February 2014].

Nosek, B.A., Smyth, F.L., Hansen, J.J., Devos, T., Linder, N.M, Ranganath, K.A., *et al*. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes, *European Review of Social Psychology*, **18**, pp. 36–88.

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

Vohs, K.D., *et al.* (2008) Making choices impairs subsequent self-control: A limited-resource account of decision-making, self-regulation, and active initiative, *Journal of Personal Social Psychology*, **94**, pp. 883–898.

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

Ziegert, J.C. & Hanges, P.J. (2005) Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias, *Journal of Applied Psychology*, **90** (3), pp. 553–562.