References

Bishop, J. 2014. Causal pluralism and the problem of natural agency. *Res Philosophica* 91: 1–10.

Davidson, D. 1984. Thought and Talk. In *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, Chapter 11.

Consciousness and Moral Responsibility

FELIPE DE BRIGARD

> *If anyone sins and does what is forbidden in any of the Lord's commands, even though they do not know it, they are guilty and will be held responsible.*
>
> Leviticus 5:17.

On 18 June 2014, Justin Ross Harris found the lifeless body of his 22-month-old son inside his car, which he left parked all day at more than 92°. Although he claims to have been completely unaware of the fact that his son was in the car, many think Harris is still responsible for his son's death, and he now faces up to 20 years in jail. These kinds of cases, in which an agent is unaware of an action – or of the reasons behind an action – but nonetheless is held responsible for it, have motivated several philosophers to reject *the Consciousness Thesis* (CT): 'the thesis that consciousness of at least some of the facts that give our actions or omissions their moral significance is a necessary condition of moral responsibility' (vii). Neil Levy disagrees, and in the six chapters that compose *Consciousness and Moral Responsibility* he mounts a clear, forceful and compelling argument to defend CT.[1] One of the many virtues of this impressive volume is its brevity. For that reason, many of Levy's assertions are underdeveloped. My hope in this note is to offer a number of observations pertaining empirical claims that could be refined and revised, without necessarily altering the gist of the argument, to improve upon what is already a persuasive view.

Levy begins, in Chapter 1, by summarizing two kinds of arguments that have motivated the denial of CT. On the one hand, some arguments deny CT based on scientific evidence suggesting that we are often unconscious of the reasons why we act but are nonetheless held responsible. On the other hand, some arguments deny CT based on the description of ordinary cases and folk practices, which allegedly provide philosophical reasons to deny CT. Although brief, the discussion here does a good job explaining the grounds upon which some philosophers deny CT. But what are these philosophers denying?

1 *Consciousness and Moral Responsibility*. By Neil Levy. Oxford: Oxford University Press, 2014. xvi + 158 pp. £27.50.

Analysis Reviews Vol 75 | Number 4 | October 2015 | pp. 661–667 doi:10.1093/analys/anv012

The answer comes in Chapter 2, where Levy distinguishes two ways to understand the claim that CT is false. One way is to understand it as the conclusion of this argument:

(1) Consciousness does not play a causal role in bringing about an action.
(2) Because consciousness does not play a causal role in bringing about the action an agent is held responsible for, then it cannot be a necessary condition for moral responsibility.
(3) Therefore, consciousness is not a necessary condition for moral responsibility.

To support (1), philosophers have made use of well-known experimental results (e.g. Libet el al. 1983; Wegner 2002), which allegedly show that the conscious decision to initiate an action is epiphenomenal. Levy does not think that the deniers of CT have this sort of reasoning in mind, and he rejects the alleged import of these results for the discussion on moral responsibility. He shows that either the empirical evidence says nothing about the protracted deliberative process that causally precedes the actions people are normally held responsible for; or that the notion of consciousness at play is confined to such an infinitesimally short mental state – i.e. the precise millisecond in which a decision is made – that any CT limited to such a notion of consciousness would not just be false, but also utterly inconsequential for moral responsibility.

Instead, Levy advocates for a different reading of the claim that CT is false. He believes that what threatens moral responsibility is the denial of the claim that we have to be conscious *of the deliberations* that bring about a particular action (24). Accordingly, the deniers of CT must have in mind a different notion of 'consciousness'. Levy distinguishes two senses in which one can be conscious of a mental content: *phenomenally* (i.e. there is something that it is like to be in a mental state with such content) and *informationally*. One is informationally conscious or *aware* of mental contents that are personally (versus sub-personally) available to the agent to report or reason with, and a content is personally available 'when the agent is able to effortlessly and easily retrieve it for use in reasoning and it is online' (33). Levy thinks that awareness – as opposed to phenomenal consciousness – is the concept of consciousness at stake in the CT.

Levy accepts that his definition of 'personal availability' is partially speculative. Still, he insists that 'only when information is available for easy and effortless recall and also online is it available to underwrite justified attributions of moral responsibility' (33). This way of understanding 'personal availability' invites us to treat it as a dichotomous process (e.g. information is either personally available or it is not), but personal availability depends upon a process (i.e. recall effort) that is more likely continuous. Recall is neither totally effortful nor completely effortless; it comes in degrees, in part, because there are a number of factors that affect retrieval, many of which vary along continuous dimensions. One such factor is retrieval support. Attempts to recall declarative information, such as episodes or facts, occur in the presence of endogenous or exogenous cues. Often, there is a large amount of overlap between the cue and the memory trace, so retrieval is effortless; this is true, for instance, when the studied item is fully reinstated during retrieval, as the task consists simply in determining whether or not such an item has been studied before. Other times the cue offers only partial information of the stored item, such as its colour, category or location. Depending on the number and/or relevance of these cues, the

amount of effort required to recall stored information varies. Sometimes the cue offers no information about the stored memory, as in free-recall tasks where individuals are asked to remember as many items as possible from a study session. Under these conditions, recall is maximally effortful.

These results were summarized by Endel Tulving (1983) in his model of 'synergistic ecphory', which he understood as the process in which a cue activates the right memory trace to bring about a successful recollection. According to this model, retrieval effort is inversely proportional to the amount of retrieval support – a continuous variable. A consequence of this model is that the same information can be made personally available with varying degrees of effort, depending on the amount and kind of retrieval support. But Levy doesn't want to treat any information that can be possibly retrieved given the right amount of support as personally available. He wants information to count as personally available only when it can be recalled 'easily and effortlessly' when there is 'no special prompting', i.e. when there is a minimal amount of retrieval support (34). But this is precisely the condition under which most retrieval effort is required. So what gives?

I think Levy mistakenly equates an elaborate cue with a cue that provides more retrieval support. A cue can be elaborate and provide no retrieval support at all, or it can be simple and sufficient to reactivate a memory trace. But if Tulving's model is correct – and the amount of data it fits gives us no reason to doubt it – the amount of effort it takes to make stored information personally available is usually not up to the subject. Consider the *encoding specificity principle*, according to which a person is more likely to retrieve an item if the context of retrieval reinstates that of its encoding (Morris et al. 1977). Under conditions of full reinstatement, what allows the subject to put less effort into retrieval has nothing to do with the person, but with the context in which the person is attempting to recall the information. Consequently, a change in context can affect whether retrieved information is personally available.

Another continuous dimension upon which retrieval effort varies is non-pathological aging. As we age, we need more and better retrieval support to bring to mind the same information that, years ago, would have required less support to recall (cf. Hedden and Gabrieli 2004). But memory decline in aging is gradual. And it would seem arbitrary to say that an informational item that is retrieved at the age of 25 with little retrieval support counts as personally available when the same item, retrieved 50 years later with more retrieval support, does not. Maybe this is what Levy wants to say. Maybe he believes that personal availability hinges on a number of factors that do not depend on the subject (Levy 2011), such as age and context of retrieval, and that such factors should influence our assessments as to whether or not the subject was aware of the relevant facts and, consequently, whether or not the subject is morally responsible for her actions. By building the notion of 'personal availability' on factors that load upon continuous dimensions, Levy may be inadvertently committed to the possibility that moral responsibility may also come in degrees. He does not explore these consequences, but they are exciting avenues for further research.

At any rate, once Levy has settled on the claim that the notion of consciousness at stake in CT is 'awareness', he moves on, in Chapter 3, to evaluate the role that awareness plays in our cognitive economy. He subscribes to the global neuronal workspace (GNW) hypothesis (Baars 1988; Dehaene and Naccache 2001), which suggests that mental contents become conscious when they are amplified and

broadcast to distinct regions of the brain, so that information is effectively processed by a global (as opposed to local) modality-independent (as opposed to modality-specific) neuronal network. As far as it goes, the GNW hypothesis is one of the best ones in the offing, despite some apparent counter-evidence (see Prinz 2012: 30–31). However, I worry that his two main reasons for taking the GNW hypothesis to be an attractive theory are *not* unique to the architecture of the GNW.

Levy follows Carruthers (2006) in believing that the brain is massively modular and that it 'does not seem well designed for domain-general information processing' (41). But since the version of CT he defends holds that what is required for moral responsibility is awareness *of the deliberations* that bring about our actions, and because deliberation is a domain-general and modality-independent information process (40), he needs the cognitive/neural mechanisms underlying awareness to be capable of processing domain-general and modality-independent information. This gives him a second reason to vouch for the GNW, as the neural architecture associated with it (which largely overlaps with the default network (see later)) integrates across modalities (Dehaene and Changeaux 2011), which in turn could strip the integrated information off its domain-specificity, endowing it – presumably – with domain-generality.

But it seems to me that current evidence from cognitive neuroscience suggests that both of Levy's reasons to side with the GNW are likely false: the brain is probably *not* massively modular, and cross-modal integration is *not* unique to the neural network associated with the GNW. Empirical evidence and theoretical arguments against the massive modularity hypothesis abound (e.g. Samuels 2006; Wilson 2008). It has been suggested that talk of domain-specific modules obfuscates the fact that domain-specificity is usually level- and process-dependent; at a certain level of description and for a certain process a brain mechanism may be considered modular, but at another level and for another process it may not. Moreover, there is evidence that most brain regions – if not all of them – are constantly redeployed at different stages within and among different cognitive processes (Anderson 2010). Finally, recent studies have shown that the rapid evolutionary expansion of the cerebral cortex in humans, relative to other primates, occurred precisely in brain areas involved in association networks; that is, in brain regions that underwrite processes of integration of information from different modalities (Buckner and Krienen 2013). These facts, along with numerous other arguments, suggest that the brain is not massively modular, that the kinds of operations for which modularity may still be predicated are few and far between (Prinz 2006), and that, in fact, the brain seems well-equipped to deal with the kind of sensory-integration and cross-modality Levy finds essential for complex information processing.

This last point dovetails with a second line of evidence for the claim that the neural architecture associated with GNW is *not* the only one that supports cross-modal integration and domain-generality processing. Using resting-state functional connectivity MRI data, researchers are now able to map large-scale functional networks in the human brain. Using a variety of analyses – such as hierarchical clustering, which determines the minimal amount of maximally coherent brain regions that are functionally co-active – a number of studies suggest that functional brain connectivity involves at least seven basic and relatively independent functional neural networks (Power et al. 2011; Yeo et al. 2011), of which at least four integrate cross-modal information from different domains (i.e. default, dorsal attention, ventral attention

and fronto-parietal task control networks). The neural network associated with the GNW appears to be one of several networks that process cross-modal information. So why should we prefer it over all the other functionally connected networks which also integrate disparate information across modalities and underwrite a number of overt behaviours? A possible answer is phenomenology, as only the network associated with the GNW has been linked to phenomenal consciousness. But this won't do for Levy, because he has already argued that phenomenology is not the relevant concept of consciousness at stake in CT.

Maybe the answer comes in Chapter 4, where Levy suggests that the kind of integration that matters for moral responsibility is the flexible combination of reason-responsive personal-level attitudes that play a role in action in a non-script manner. It is possible that only the functional network associated with GNW can afford 'genuine flexibility of response' – that is, 'sensitivity to the content of a broad range of cues' (76) – whilst the others may only process 'scripts', which are simply responsive 'to a certain overlearned range of cues' (76). Ultimately this is an empirical hypothesis, but it is an attractive one, despite the fact that it relies on Carruthers's view of cognitive flexibility, which is not without controversy (Machery 2008; Weiskopf 2014).

The difference between actions brought about as a result of the flexible combination of contents (i.e. contents one can be aware of) versus those that are only the effect of script-like associations, plays a critical role in the rest of the book. In Chapter 5, Levy argues that only contents that can be genuinely flexibly combined in the GNW include personal-level attitudes that can be reason-responsive and express evaluative agency, i.e. the kinds of mental contents that express the real self. Similarly, in Chapter 6, Levy argues that only these sorts of mental contents can allow the agent to exercise responsible control. Thus, if only contents that are broadcast in the GNW include the kind of personal-level attitudes capable of grounding moral responsibility, then awareness becomes necessary for moral responsibility. And, if so, CT has been vindicated.

My only concern regarding the last two chapters of this really wonderful book, is that Levy's strategy for distinguishing personal and sub-personal level contents is not completely clear. According to his view, our evaluative agency is constituted by our coherent personal-level attitudes' (90), among which he lists 'concerns, beliefs, commitments and goals' (89), and only those contents that can be assessed for consistency with our personal-level attitudes can be broadcast. As such, only the contents of personal-level attitudes enter the GNW. But what exactly is the relation between awareness and personal-level contents? At the beginning of the book, Levy claims that 'consciousness makes information available for the rational control of behaviour by making it available for use in deliberation and reflection' (63), suggesting that consciousness *causes* certain contents to become personal level. However, in the second half of the book, it seems that consciousness does not play such a causal role, and that contents come pre-sorted as only those that already are personal level can be broadcast. If so, one wonders what properties do contents that can feature in personal-level attitudes have, so that they, unlike the script-like contents of sub-personal (implicit) attitudes, can be broadcast in the GNW?

My own answer would be that only contents that can be (internally or externally) attended to can be broadcast to the GNW (De Brigard 2011). But Levy does not consider the possibility that attention could play this causal role. Instead, he relies on a

difference in the acquisition of such contents. Specifically, he suggests that sub-personal level contents, such as implicit attitudes, are acquired by associative systems 'which respond to regularities in the environment' (98) – they are 'just associations' (105) – while, presumably, personal-level contents do not. Levy even conjectures that the dopaminergic circuit underlying the reward system may be responsible for the acquisition of some sub-personal-level attitudes, which he takes as evidence of its independence from the mechanisms underwriting the acquisition of personal-level contents. Although (as Levy himself acknowledges) this conjecture is controversial (98), I think it is more likely to come out false, because there is already enough evidence showing that dopaminergic circuits in the mid-brain and striatum play a critical role in the acquisition (Schott et al. 2006) and retrieval of declarative information (Scimeca and Badre 2012), which – as discussed earlier – constitute the contents that, upon recall, can be broadcast to the GNW. Moreover, there is also behavioural evidence showing that higher-order regularities in the environment also play a critical role in scaffolding the encoding and retrieval of declarative information (Brady and Oliva 2008).

As mentioned at the beginning, many of the issues raised in this note are a natural consequence of the fact that some of the book's claims are underdeveloped. This should not be seen as a weakness, though, but as a virtue. For unlike much philosophical work on moral psychology, Levy does not ground his empirical claims on shaky private intuitions elicited by under-described imaginary cases, but on the inter-subjectively assessable results of psychology and neuroscience, a methodological strategy that allows him to clearly demarcate avenues for further investigation. With this book, Levy has given us further reason to accept that moral psychology cannot – and should not – be done separately from the sciences of the mind. *Consciousness and Moral Responsibility* is thus not only a fabulous contribution to ethics, philosophy of mind and philosophy of action, but also a clear instance of how exemplar moral psychology ought to be done.[2]

Duke University
Durham, NC 27708, USA
felipe.debrigard@duke.edu

References

Anderson, M.L. 2010. Neural reuse: a fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33: 245–66.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

Brady, T.F. and A. Oliva. 2008. Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological Science* 19: 678–85.

Buckner, R.L. and F.M. Krienen. 2013. The evolution of distributed networks in the human brain. *Trends in Cognitive Sciences* 17: 648–65.

2 Thanks to Bryce Huebner and Walter Sinnott-Armstrong for their comments and suggestions.

Carruthers, P. 2006. *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

De Brigard, F. 2011. The role of attention in conscious recollection. *Frontiers in Psychology* 3: 1–10.

Dehaene, S. and J.P. Changeaux. 2011. Experimental and theoretical approaches to conscious processing. *Neuron* 70: 200–27.

Dehaene, S. and L. Naccache. 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79: 137.

Hedden, T. and J. Gabrieli. 2004. Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews Neuroscience* 5: 87–96.

Levy, N. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

Libet, B., C.A. Gleason, E.W. Wright, and D.K. Pearl. 1983. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* 106: 623–42.

Machery, E. 2008. Modularity and the flexibility of human cognition. *Mind and Language* 23: 263–72.

Morris, C.D., J.D. Bransford, and J.J. Franks. 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* 16: 519–33.

Power, J.D., A.L. Cohen, S.M. Nelson, G.S. Wig, K.A. Barnes, J. Church, A. Vogel, T.O. Laumann, F.M. Miezin, B.L. Schlaggar, S.E. Petersen. 2011. Functional network organization of the human brain. *Neuron* 72: 665–78.

Prinz, J. 2006. Is the mind really modular? In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 22–36. Oxford: Blackwell.

Prinz, J. 2012. *The Conscious Brain*. Oxford: Oxford University Press.

Samuels, R. 2006. Is the human mind massively modular?. In *Contemporary Debates in Cognitive Science*, ed. R. Stainton, 37–56. Blackwell: Oxford.

Schott, B.H., C.I. Seidenbecher, D.B. Fenker, C.J. Lauer, N. Bunzeck, H. Bernstein, *et al.* 2006. The dopaminergic midbrain participates in human episodic memory formation: evidence from genetic imaging. *Journal of Neuroscience* 26: 1407–17.

Scimeca, J.M. and D. Badre. 2012. Striatal contributions to declarative memory retrieval. *Neuron* 75: 380–92.

Tulving, E. 1983. *The Elements of Episodic Memory*. Oxford: Clarendon Press.

Wegner, D.M. 2002. *The Illusion of Conscious Will*. Cambridge: MIT Press.

Weiskopf, D. 2015. Modularity and cognitive architecture. In *An Introduction to Philosophy of Psychology*, eds. F. Adams and D. Weiskopf. Cambridge: Cambridge University Press.

Wilson, R.A. 2008. The drink you're having when you're not having a drink. *Mind & Language* 23: 273–83.

Yeo, B.T.T., F.M. Krienen, J. Sepulcre, M.R. Sabuncu, D. Lashkari, *et al.* 2011. The organization of the human cerebral cortex estimated by functional connectivity. *Journal of Neurophysiology* 106: 1125–65.