T

Mario De Caro, Massimo Marraffa

Consciousness and responsibility

There undoubtedly is a strong tension between cognitive science and folk psychology. On the one hand, some cognitive scientists drastically downplay introspection, and with that they cast radical doubt on the ordinary conception of ourselves as conscious agents: except for perceptual data, they claim, conscious mental states are illusionary. On the other hand, naive ethics – as reconstructed by experimental philosophy – looks to consciousness as the fundamental basis for attributing responsibility: agents are responsible for an action if it reflects a conscious deliberation on their part.

After exposing this disagreement, we will advocate adopting an intermediate position between traditional philosophers, who continues to ascribe primacy to consciousness in action in spite of the data emerging from the mind-brain sciences, and scientists (or empirically oriented philosophers) who, overgeneralizing from specific cases, claim that all conscious mental states are epiphenomenal. An example of this intermediate position can be gleaned from some authors (Levy, 2014; Carruthers, 2015a; Carruthers and King, 2022), who convincingly argue that cognitive neuroscience, rather than proving the epiphenomenalism of consciousness, allows for a finer-grained articulation of the dialectic between unconscious processing and conscious reflection.

1. *Introspection as theorizing*

In the last decades, a psychological tradition of research has developed experimentally the Freudian hypothesis of our propensity for self-decep-

 DOI: 10.4454/teoria.v42i2.158

tion, i.e. a tendency to fabricate "convenient" explanations of our conduct. This has happened especially in social and group psychology, where experimental designs have been devised with participants that have no direct introspective access to their real motivations (i.e., the true causes) of their conduct in the experiment; unaware of these motivations, they nevertheless fabricate a posteriori – on the basis of socially shared explanatory theories or idiosyncratic theorizing – reasonable but imaginary explanations of their own conduct (a form of nonclinical "confabulation"). Here, unconscious everyday mechanisms of self-deception have been shown to be more pervasive, articulate, varied, and profound than Freud thought (cf. Wegner, 2002; Wilson, 2002; Johansson *et al*. 2013).

Consider a classic case of confabulation of intentions. In a study by Wegner and Wheatley (1999), a participant P and an experimenter's accomplice rested their fingers on a tablet mounted on a computer mouse, moving a cursor on a screen where about fifty small objects appeared. Subjects heard words in headphones and had to keep moving the mouse until the stop signal came (about every 30 sec). P was induced to mistakenly believe that she was the one who made the decision to stop the cursor movement; this was achieved by having her listen to the name of one of the objects that appeared on the screen just before the accomplice locked the cursor next to the image of the named object. In addition to the confabulation of decisions, there were fluctuations in the perception of intentionality depending on when P heard the word.

These kinds of experimental data (which could be multiplied at will) are the source of theories in which "introspection" is judged to be a misnomer for an interpretive process, that is, a process that makes use of information concerning states of affairs external to the mind (the agent's manifest behavior and/or the situation in which that behavior takes place) in order to *theorize* about the causal etiology of one's own and others' behavior. This is the theory of self-knowledge that establishes a *Self/Other Parity* (cf. Schwitzgebel, 2019, §2.1), whose historical referent is Ryle (1948)[1].

In this view, introspective consciousness is redefined as the ability to ex post facto remotivate one's actions, that is, the ability to continuously "approve" what one is doing. The agent is no longer – as a stereotype

[1] "The sort of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same […I]n principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe" (Ryle, 2009, p. 139).

implicit in the naive way of examining animal-type living systems would have it – a primarily quiescent organism, which 'then' moves, each time for a given purpose; it is rather a primarily self-propelled structure. So, one can never really tell when an action begins nor when an identifiable plan of behavior directed toward an end arises. It is more accurate to say that we have always been immersed in a system of behavioral patterns (or, more precisely, cognitive-motor patterns) that we have begun to articulate since we exist as individuals, and that we relentlessly modify and repurpose according to circumstances and the stimuli that modulate them. And immersed in this flow of actions, we sometimes say and tell ourselves "This is just the thing I want to do", or "What I did is the thing I really wanted to do", or again "This thought is just what I feel like thinking". In this view, what characterizes "voluntary" human action is not so much the presence of anticipatory mental events, but (i) the fact that we are not surprised that we have performed that action[2]; and (ii) that we then explain it. As Anscombe (1957) noted, it is incorrect to assume that we know what our intentions are; what is to correct to say, rather, is that we can tell what our intentions are.

2. *Do conscious thoughts exist?*

Having reached this point, it is important to note that no serious scholar has endorsed a *purely* self/other parity view. Nisbett and Wilson (1977), for example, distinguished between "cognitive processes" (i.e., the causal processes underlying judgments, decisions, emotions, and feelings) and mental "content" (the judgments, decisions, emotions, and feelings themselves). This private content can be accessed directly, resulting in knowledge endowed with "almost complete certainty". And Ryle (1949) himself, when he stresses the importance of outward behavior in our mentalistic self-attribution practices, acknowledges the presence of "twinges", "thrills", "tickles", and even "silent soliloquies", which we know of in our own case and that do not appear to be detectable by observing outward behavior. However, since none of these scholars has offered any hypothesis about the mechanisms of this apparently more direct self-knowledge, their

[2] "[D]ie willkürliche Bewegung sei durch die Abwesenheit des Staunens charakterisiert" ("Voluntary movement is marked by the absence of surprise") (Wittgenstein, 1953, Engl. transl. 1986, §628).

theory is *incomplete* (Schwitzgebel, 2019, §2.1). With this in mind, it is of the utmost importance to turn attention to Peter Carruthers's (2011, 2015a, 2019) enhanced version of the self/other parity view.

Carruthers's theory of introspective self-knowledge rests on the validity of a global workspace account of the conscious accessibility of our perceptual experiences, first postulated by Baars (1988) and widely confirmed since (Dehaene, 2014). In particular, analyses of functional connectivity patterns in the human brain have shown which sort of neural architecture is necessary to realize the main elements of a global broadcasting account. Specifically, these studies show the existence of two main neurocomputational spaces within the brain, each characterized by a distinct pattern of connectivity.

The first space is a processing network, composed of a set of parallel, distributed, and functionally specialized processors or modular subsystems subsumed by topologically distinct cortical domains with highly specific local or medium-range connections that encapsulate information relevant to its function. These subsystems compete with each other to access the Global Neuronal Workspace (GNW), which is implemented by long-range cortico-cortical connections, mostly originating from the pyramidal cells of layers 2 and 3 that are particularly dense in prefrontal, parieto-temporal and cingulate associative cortices, together with their thalamo-cortical loops.

The global broadcasting architecture provides Carruthers with a framework within which it can be argued that *occurrent thoughts* are always unconscious and direct the stream of consciousness and reflection from behind the scenes. The expression "occurrent thoughts" refers to propositional attitude events (such as "judging something to be the case", "deciding to do something", or "actively intending to do something") that are *episodic* rather than persisting, and have a *non-sensory format* (they are "amodal"). Carruthers claims that only sensory or sensory-involving states can participate in consciousness (and, a fortiori, reflection), while amodal propositional attitudes operate unconsciously in the background. This thesis is argued in two steps.

First, according to Carruthers occurrent thoughts cannot be *first-order* access-conscious. The global broadcasting architecture affords to explain the conscious accessibility of our sensory or sensory-involving states. When one of the functionally specialized processors accesses the global workspace, its outputs (i.e., sensory information including perceptions of the world, the deliverances of somatosensory systems, imagery, and inner

speech) are broadcast to an array of executive, conceptual, and affective "consumer" systems. These systems process ("consume") sensory information according to their various specialisms – e.g., drawing inferences, forming memories, producing emotional responses, forming judgments, planning and making decisions, and verbally reporting. By contrast, thoughts – that is, the outputs of the consumer systems – are not capable of being globally broadcast. The reason is that the mechanism by which a state is broadcast is *top-down attention*; and in reviewing the literature on attention in cognitive neuroscience, Carruthers finds that "attention itself has an exclusively sensory focus", primarily targeting "midlevel sensory areas" (2015a, pp. 91-2). (More precisely, a top-down attentional network links the dorsolateral prefrontal cortex, the frontal eye-fields, and the intraparietal sulcus. The "business end" of the system is the latter, which projects both boosting and suppressing signals to targeted areas of mid-level sensory cortices.) Hence the anticipated conclusion: only states with a sensory-based format are capable of becoming first-order access-conscious.

Let us come to Carruthers's second argumentative step. According to him, occurrent thoughts cannot be *higher-order* access-conscious either. It seems obvious that thoughts are available in a way that enables us to know of their occurrence without requiring self-interpretation, of the sort that makes us aware of the thoughts of other people. The global broadcasting architecture, however, allows Carruthers (2011) to develop a robust version of the self/other parity account of self-knowledge. According to Carruthers' version of the self-other parity theory of the nature and sources of self-knowledge (the so-called "Interpretive Sensory-Access", ISA), we can have non-interpretive access only to our sensory or sensory-involving states; all knowledge of our own occurrent thoughts is instead a matter of *interpretation*.

Among the consumer systems that form judgments (i.e. events of belief-formation), a "mindreading system" exists that is a multi-componential faculty that exploits a corpus of folk-psychological theoretical knowledge in order to generate metarepresentational beliefs about the mental states of others and of oneself. This faculty, Carruthers argues, was originally designed for "reading" other minds; only at a later stage the ancestral mindreaders started to apply this skill to themselves, forming beliefs about their own mental states as they did about other people's. Since the mindreading system evolved for understanding other people, it is *outward looking*: it has access to all sensory information broadcast by our perceptu-

al systems, and hence it also has non-interpretive access to one's own sensory states. However, it does not give us direct access to our own thoughts; so we must infer them from observations of our circumstances and behavior, interpreting ourselves just as we interpret others. In this light, the only difference between self- and other- knowledge of thoughts is that in one's own case, the mindreading system has more available information upon which to base its interpretation. As a matter of fact, in addition to using overt behavior, in one's own case it can also draw on a subject's affective, sensory, and quasi-sensory states such as visual imagery or inner speech tokens that are globally broadcast in the mind. In brief, Carruthers's ISA theory restricts self/other parity to a particular subclass of mental states, i.e. propositional attitude events as opposed to mental events with a sensory-based format, which are introspectable (cf. Schwitzgebel, 2019, §§ 2.1.3 and 4.2.2).

Here, then, is how the ISA theory is able to explain what earlier versions of the self/other parity failed to explain, namely, why mentalistic self-attribution can occur even in the absence of behavioral and contextual data, and why one is able to "read" one's own mind better than that of others. Even when I am sitting in my room, motionless and with my eyes closed, I have no difficulty in attributing mental states to myself because I can still rely on a great deal of information regarding the situation I am in, in the form of sensory, imaginative and somatosensory data.

The moral to be drawn is an eliminativist in relation to conscious thought. Since the distinctive feature of the global-broadcasting mechanism is that it is sensory-based, amodal propositional attitudes cannot broadcast themselves, though they might cause sensory-like events (e.g., a sentence in inner speech) which are so broadcast. Outside of the broadly sensory domain (sensation, perception and affect) none of our mental states is ever conscious.

The disappearance of conscious thought still leaves room for a distinction between unconscious *intuitive* processes and conscious *reflective* processes. The latter are forms of mental activity that are directed, for example, toward solving a problem, arriving at a judgment, or reaching a decision. These reflective processes rest on *working memory*, the executive system for directing attention and sustaining and manipulating imagery in the global workspace; and working memory is a *sensory-based* system. First, working memory is a process that emerges and constitutively depends on sensory systems (Postle, 2006); second, top-down attention directed at mid-level perceptual regions of the brain is necessary not only for

conscious perception but also for that contents to enter working memory. The latter uses top-down attention to activate and sustain imagistic representations in conscious form; there is no place within it for amodal propositional attitudes. Since working memory is the system that underlies conscious reflective processes, the latter must be sensorily laden. Supposed conscious thoughts are sensory images in working memory, typically imaged utterances[3].

It is of utmost importance to note that within this framework consciousness is by no means an epiphenomenon, since it performs an essential coordinating function in the mental lives of humans and many other animal species. Perceptual information becomes available to consumer systems only by virtue of global diffusion, and this allows them (and thereby the entire organism) to coordinate around a "common focus"[4].

Even so, the essential feature of the global broadcasting mechanism is its sensory character: an amodal propositional attitude event cannot be globally broadcast, although it can cause a sensory event that can be (e.g., a sentence in internal language). So, except for the sensory domain (sensations, perceptions, and emotions), none of our mental states is available to access consciousness. In particular, there are no entities such as (nonperceptual) judgments, intentions or conscious decisions.

3. *The nexus of moral responsibility and conscious thought in naive ethics*

If ISA theory is well grounded, it puts a strong constraint on the construction of a theory of (moral and legal) responsibility congruent with the findings of neurocognitive sciences: the existence of conscious amodal

[3] As Gomez-Lavin (2017) noted, Carruthers' philosophical treatment of the constructs of attention and working memory leads us to Aristotle's *De Anima*, where the capacity of *phantasia*, like working memory, enables us to entertain a perceptual image in the absence of any stimulus; more crucially, *phantasia* is deemed necessary for all thought, as "the soul never thinks without an image" (431a16).

[4] "Consciousness does make a difference. Indeed, it is vital to the overall functioning of the human mind. […] I certainly don't think consciousness is epiphenomenal. On the contrary, it plays a crucial coordinating function in the minds of humans and most other animals. It is only when information becomes globally broadcast (= becomes access-conscious) that it is made available to a wide range of down-stream systems for drawing inferences, forming memories, evaluating, and so on. This enables all those systems (and thereby the organism as a whole) to become coordinated around a common focus." (Carruthers, 2015b, 1 e 7 agosto).

propositional attitude events cannot be among the theory's commitments (King and Carruthers, 2012, 2022).

That naive ethics establishes a link between moral responsibility and conscious intentional mental states seems to be attested by some research conducted in the field of experimental philosophy applied to the concepts of freedom and responsibility. In the free will debate, philosophers often resort to ordinary intuitions – in particular, it is often claimed that naive ethics is *incompatibilist*. However, Nahmias, Morris, Nadelhoffer and Turner (2006) have argued that their experimental results attest to precisely the opposite: common sense is – as Strawson (1962) had already argued – *compatibilist*. However, Nichols and Knobe (2007), reviewing the findings of Nahmias' group, wondered why so many philosophers who are interested in the question of free will today have become convinced of the incompatibilist nature of ordinary intuitions. Their hypothesis is this: there may be a tendency in people to provide compatibilist answers to concrete questions about particular cases, but incompatibilist answers to abstract questions about general moral principles. If so, the divergence between the data of psychological studies and the conclusions of philosophers would be attributable to a difference between two different ways of *framing* the relevant question.

To test this hypothesis, Nichols and Knobe presented participants with descriptions of two universes, A and B. Universe A is a universe in which everything takes place in accordance with deterministic laws. In universe B, on the other hand, everything occurs in accordance with deterministic laws except for human decisions. Participants were first asked the question "Which universe is most similar to ours?" to which 90% responded by opting for the indeterministic universe B. Then participants were randomly assigned to one of two conditions, abstract and concrete.

Participants placed in the *abstract* condition were asked the following low-emotion question: "In universe A is it possible for a person to have full moral responsibility for his or her actions?" In this condition 86% of the participants gave the incompatibilist answer that in universe A full moral responsibility is not possible. In contrast, participants in the *concrete* condition were presented with a deterministic universe in which a specific agent, Bill, committed a morally reprehensible act (killing his wife and children). The question was: "In your opinion, does Bill bear full moral responsibility for the death of his wife and children?" (Nichols and Knobe, 2007, p. 670). In this concrete and emotionally charged condition, 72% of the subjects gave the compatibilist response that Bill bears full moral responsibility for the murder of his wife and children.

Thus, these data seem to confirm the hypothesis that intuitions about the determinism/responsibility relationship vary depending on the emotional framing of the imagined case. When participants are confronted with macroscopic violations of moral norms, they experience a strong affective reaction (a *reactive* attitude such as moral anger or indignation) that renders them unable to properly apply the underlying naive theory of moral responsibility, which – Nichols and Knobe argue – is incompatibilist. Compatibilist intuitions are then the result of a *performance error* caused by the disruptive influence of emotion on moral judgment. In other words, the bias triggered by strong affect prevents subjects from making the inference that is instead made at the abstract level, leading to the conclusion that determinism excludes responsibility. From this perspective, the conclusion is that the compatibilist intuitions of the ordinary individual are only apparent; and must be set aside as they are subject to the distorting influence of emotional responses.

According to Eddie Nahmias and collaborators (Nahmias, Coates and Kvaran, 2007; Nahmias and Murray, 2010; Nahmias, 2011), however, the scenarios constructed by Nichols and Knobe do not allow the results of their study to be interpreted as evidence of the incompatibilist character of the naive theory of moral responsibility. In fact, Nahmias *et al.* argue, what led the participants in the experiment to deny free will and moral responsibility is the interpretation of determinism as a thesis that implies the idea that the causes of behavior *bypass* the conscious and rational control of the agent. In other words, the description of determinism used by Nichols and Knobe ("everything must occur the way it does in fact occur") may have suggested to the participants that conscious deliberations and ends play no causal role in determining the agent's conduct – i.e. they are *epiphenomenal*[5]. And, indeed, if determinism is interpreted in terms of "bypassing" consciousness, compatibilism is actually doomed (since this view implies that conscious mental states play a relevant role in the generation of action); and the same happens if determinism is interpreted as a form of fatalism – that is, as the belief that certain events will take place regardless of what we decide or try to do. However, Nahmias et al. maintain, determinism does *not* entail bypassing or epiphenomenalism about mental

[5] The key difference is that in universe A each decision is completely caused by what happened before the decision – given the past, each decision must be made the way it is in fact made; in universe B, on the other hand, decisions are not completely caused by the past, and each decision does not have to be made the way it is in fact made.

states or fatalism. In a deterministic universe, natural events remain contingent. Also, determinism does not exclude that conscious mental states play a causal role in human conduct. Quite the contrary: to the extent that our mental states are part of a deterministic sequence of events, they play an essential role in determining what will happen. On this view, then, it is not so much that freedom and responsibility are threatened by determinism as such, but only when it is conceived of as a *reductionist mechanism* – that is, when it is claimed that the higher-level properties of a system (and its changes over time) are reduced to and can be exhaustively explained by its lower-level mechanisms – as when human conduct is reduced to the causal mechanisms of the nervous system in which conscious mental states play no role. In brief, reductionist mechanism asserts that human actions are caused by lower-level mechanisms rather than by his conscious mental states and rational capacities.

In short, while Nichols and Knobe argue that judgments made in high-emotional-impact cases are the outcome of a performance error attributable to the disruptive influence of our emotions and from this conclude that the naive concept of responsibility is incompatible with the truth of determinism, Nahmias *et al*. advance the opposite thesis, namely that performance errors take place when participants mistakenly assume that determinism excludes the possibility of conscious, rational control.

4. *Reconceptualizing the consciousness thesis*

As said, there is some evidence that naive ethics looks to consciousness as the fundamental basis for attributing responsibility – an agent is responsible for an action if it reflects a conscious deliberation on his part. This was translated into normative terms by Levy (2014), who argued for 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" (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. Furthermore, he argues that the two leading accounts of moral responsibility – *real self* account (Frankfurt, 1971, 1988) and *control-based* account – are committed to the truth of the consciousness thesis despite what proponents of these accounts maintain. According to Levy, (a) only the actions that are performed consciously can express our evaluative agency, and the expression of moral attitudes requires consciousness of that attitude; and (b) we possess responsibility-level control only over actions that we perform consciously, and control over their moral significance requires consciousness.

However, the consciousness thesis seems to contradict the constraint that ISA theory imposes on the construction of a theory of responsibility. In fact, to be congruent with data from the neurocognitive sciences, such a theory must not presuppose the existence of conscious amodal propositional attitude events. In the case of the real self, it is claimed that an agent can be held responsible exclusively for those actions that have been caused by psychological states reflecting its identity as practical agent. But if the propositional attitudes that define the agent's real self are the conscious ones, the elimination of conscious thought implies the non-existence of the real self (King and Carruthers, 2012, pp. 217ff).

Now let us ask: would a theory of responsibility that satisfies this constraint allow us to preserve at least part of the considerations that motivate the idea that the actions for which we are responsible are the actions that originate from conscious attitudes and decisions? For example, would such a theory allow us to distinguish between actions that originate from so-called "implicit attitudes" and actions that arise from conscious reflection? Suppose, for example, that an individual is totally unaware that they have an *implicit bias* against people of colour. Consequently, as they review some job applications, they prefer a less qualified white applicant to a black applicant. Should this person be blamed for doing so? Certainly we should be in a position to say that this individual is far less culpable than someone who, while reading the resume, thinks "I would never hire a person of colour" and for that very reason trashes the application.

According to Carruthers (2015a, §§3.3 and 3.5) this distinction can still be drawn in his ISA theory. Indeed, although the latter does not allow a distinction to be drawn between conscious and unconscious amodal attitudes (amodal attitudes being all unconscious), a kindred distinction can still be drawn – that is, one can still distinguish between attitudes that are formed by virtue of one's conscious reflections and those that are caused by unconscious processes. Attitudes that originate from conscious reflec-

tion are still unconscious attitudes (they are those whose existence is often known to the subject as the result of the mentalistic interpretation of the sensory contents of reflection). Nevertheless, they are attitudes to the formation of which the whole person has contributed (and note that here Carruthers is following Levy, 2014):

> Asking oneself in inner speech, "What should I decide?," for example, issues in a globally broadcast request for information, thereby allowing all the different consumer subsystems that receive such broadcasts a chance to contribute an answer. There is a good sense, then, in which attitudes that are formed as a result of conscious reflection are owned by the whole person, in a way that a decision to redirect attention to the sound of one's own name is not (Carruthers, 2015a, p. 237).

Within this framework, the distinction between personal and subpersonal attitudes can be reformulated. They are attitudes of the same type but differ with regard to their etiologies: personal attitudes, but not subpersonal attitudes, are unconscious attitudes that are caused by conscious reflection.

Applying this distinction to the case of the implicit bias, we obtain the following. A decision that arises from conscious reflection on the alleged demerits of people of color is one to which the whole person contributes. It therefore reflects, in a sense, *the self as a whole*. In contrast, where the decision is caused by an unconscious bias, it reflects that bias and nothing more. All of the person's other purposes and values might tend in the opposite direction, so that if his attention had been focused on the difference in competence between the two candidates as well as the implicit bias, they would have immediately chosen the black candidate.

From this perspective, cognitive neuroscience by no means leads to the epiphenomenalism of consciousness; rather, it allows for a finer-grained articulation of the dialectic between unconscious processing and conscious reflection. And this undeniably is an important piece in a theory of responsibility that aspires to hinge the normative plane on the descriptive one.

References

Anscombe, E. (1957). *Intention*. Oxford: Blackwell.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford: Oxford University Press.

Carruthers, P. (2015a). *The Centered Mind*. Oxford: Oxford University Press.

Carruthers, P. (2015b). Who's in charge anyway? Pubblicato sul blog della Oxford University Press il 1 agosto 2015: <http://blog.oup.com/2015/08/whos-in-charge-conscious-mind/>.

Carruthers, P. (2019). *Human and Animal Minds*. Oxford: Oxford University Press.

Dehaene, S. (2014). *Consciousness and the Brain*. New York: Viking.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1), pp. 5-20.

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Johannson, P., Hall, L., Tärning, B., Sikström, S., Chater, N. (2013). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It!. *Journal of Behavioral Decision Making*, https://doi.org/10.1002/bdm.1807Legal Information Institute of Cornell University.

King, M., Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy*, 9, pp. 200-28.

King, M., Carruthers, P. (2022). Responsibility and consciousness. In D. Nelkin and D. Pereboom (eds.), *Handbook of Moral Responsibility*. Oxford: Oxford University Press, pp. 448-67.

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

Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In *The Oxford Handbook on Free Will*. Oxford: Oxford University Press, 2nd edn., pp. 555-75.

Nahmias, E., Coates, D., Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy*, 31, pp. 214-42.

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

Nahmias E, Murray D. (2010). Experimental philosophy on free will: an error theory for incompatibilist intuitions. In *New Waves in Philosophy of Action*. New York: Palgrave-Macmillan, pp. 189-215.

Nichols, S., Knobe, J. (2007). Moral responsibility and determinism. *Nous*, 41, pp. 663-85.

Nisbett, R., Wilson, T.D. (1977). *Telling more than we can know: Verbal reports on mental processes*. *Psychological Review*, 84, pp. 231-59.

Ryle, G. (1949). *The Concept of Mind*. London: Routledge, 2009.

Schwitzgebel, E. (2019). Introspection. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/win2019/entries/introspection/>.

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

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

Wegner, D.M., Wheatley, T.P. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, pp. 480-92.

Wilson, T.D. (2002). *Strangers to Ourselves*. Cambridge (MA): Harvard University Press.

Wittgenstein, L. (1953). *Philosophische Untersuchungen*. Oxford: Blackwell (Engl. transl. 1986).

Abstract

Nowadays there is a strong tension between cognitive neuroscience and many ethical views based on the ordinary view of the world. On the one hand, many cognitive neuroscientists and empirically oriented philosophers raise a radical doubt about the ordinary conception of ourselves as conscious thinking agents who causally control their actions – where conscious thinking includes our beliefs, goals, decisions, and intentions. On the other hand, many ethicists still accept the ordinary conception of ourselves and, consequently, look at consciousness as one of the two fundamental bases for attributing responsibility: agents are responsible for their actions as long as such actions reflect their conscious deliberations (the other basis for the attribution of responsibility is that conscious deliberations do contribute causally to the generation of actions).

After exposing this disagreement, we will advocate the adoption of an intermediate position between that advocated by traditional ethicists (who, in spite of the data emerging from mind and brain sciences, keep attributing an absolute primacy to conscious thought in moral agency) and that held by cognitive neuroscientists and philosophers (who venture to claim that the conscious mind is indeed epiphenomenal). We will argue that an alternative and more promising model may be built by referring to some suggestions by Neil Levy, Peter Carruthers, and Matt King. In this light, we will claim that

cognitive neuroscience's findings – rather than showing that the conscious mind is epiphenomenal – require that we offer a finer-grained and unbiased articulation of the dialectic between unconscious processing and conscious reflection.

Keywords: conscious thought; experimental moral philosophy; moral responsibility; personal and subpersonal attitudes.

Mario De Caro
Università di Roma 3
mario.decaro@tlc.uniroma3.it

Massimo Marraffa
Università di Roma 3
massimo.marraffa@uniroma3.it