BRILL

Journal of Moral Philosophy 9 (2012) 200–228

JOURNAL OF MORAL PHILOSOPHY

brill.nl/jmp

Moral Responsibility and Consciousness

Matt King
St. Bonaventure University
mail.mattking@gmail.com

Peter Carruthers*
University of Maryland
pcarruth@umd.edu

Abstract
Our goal in this paper is to raise a general question about the relationship between theories of responsibility, on the one hand, and a commitment to conscious attitudes, on the other. The evidence from cognitive science suggests that there are no conscious mental states playing the right causal roles to count as decisions, judgments, or evaluations. We propose that all theorists should determine whether their theories (or the examples that motivate them) could survive the discovery that there are no conscious states of these kinds. Since we take it that theories of moral responsibility should, in general, operate with the weakest possible empirical assumptions about the natural world, such theories should be framed in such a way as to be free of any commitment to the existence of conscious attitudes, given the very real possibility that there might turn out not to be any.

Keywords
consciousness; moral responsibility; psychology; Real Self views

Our aim in this paper is to raise a question about the relationship between theories of responsibility, on the one hand, and a commitment to conscious attitudes, on the other. Our question has rarely been raised previously. Among those who believe in the reality of human freedom, compatibilists have traditionally devoted their energies to providing an account that can avoid any commitment to the falsity of determinism while successfully accommodating a range of intuitive examples. Libertarians, in contrast, have aimed to show that either physical indeterminacy or a certain kind of agent causation can find a place in the world for what they take to be

* The ordering of names was determined by a coin toss.

DOI 10.1163/174552412X625682

genuine freedom. Few have considered whether moral responsibility requires a commitment to conscious attitudes.[1]

Our question derives from a confluence of two sources. First, there is reason to think that conscious attitudes matter to theories of responsibility, either directly, as a result of the latter's commitments, or indirectly, by virtue of the assumptions that they make about certain intuitive examples. Second, there is accumulating evidence suggesting that there aren't any conscious mental states possessing the sorts of causal roles required of propositional attitudes. Since theorists of responsibility should in general be concerned to make their views compatible with plausible claims about the natural world, the implications of this data should be carefully considered. Our aim is therefore to motivate and begin exploring answers to the following conditional question: If it should turn out that there are no conscious attitudes, then what would be the implications of this fact (if any) for theories of responsibility?

We propose that theorists who aren't skeptics about moral responsibility should examine their accounts, asking whether their theories (or the examples that motivate them) could survive the discovery that there are no conscious judgments, decisions, or evaluations. Since we take it that moral theorizing in general should operate with the weakest possible empirical assumptions about the natural world, such theorists should consider whether their accounts could be motivated in such a way as to be free of any commitment to the existence of conscious attitudes. This is because, we will suggest, there is a very real possibility that there might turn out not to be any. Although the question we raise is quite general, for the most part

---

[1] One of the only examples known to us is Neil Levy, 'Are zombies responsible? The Role of Consciousness in Moral Responsibility', (forthcoming). He argues that agents are only fully responsible for their actions when those actions are a product of conscious reasoning and/or decision making. In a somewhat different spirit, philosophers like Mele and Nahmias have begun to critique the work of cognitive scientists such as Libet and Wegner, who challenge the existence of conscious will. See A. Mele, 'Decision, Intentions, Urges, and Free Will: Why Libet Has Not Shown What He Says He Has', in J. Campbell, M. O'Rourke, and D. Shier (eds.), *Explanation and Causation* (Cambridge, MA: MIT Press, 2005); E. Nahmias, 'When Consciousness Matters: A Critical Review of Daniel Wegner's *The Illusion of Conscious Will*', *Philosophical Psychology* 15 (2002), pp. 527-41; B. Libet, 'Unconscious Cerrebral Initiative and the Role of Conscious Will in Voluntary Action', *Behavioral and Brain Sciences* 8 (1985), pp. 529-66, and B. Libet, 'Consciousness, Free Action and the Brain', *Journal of Consciousness Studies* 8 (2001), pp. 59-65; D. Wegner, *The Illusion of Conscious Will* (Cambridge, MA: MIT Press, 2002). But for the most part philosophers have failed to take up the question of what, if anything, would follow for theories of responsibility if such claims made by cognitive scientists were correct. This is our question.

we will conduct our discussion within the framework of compatibilist theories of responsibility in particular. This is because we are convinced compatibilists ourselves, but also so that we can keep the discussion within manageable bounds.[2]

We begin, in Section 1, by providing some reasons for thinking that our pretheoretical beliefs about moral responsibility presuppose the existence of conscious attitudes. Then in Section 2 we briefly discuss the emerging case for denying the existence of such attitudes. (In the present context we aim to say just enough to motivate the ensuing discussion. The real debate over conscious attitudes will need to be joined elsewhere, of course.) In Section 3 we pursue the implications of such a position for one class of accounts, namely so-called "Real Self" theories of moral responsibility.[3] Our goal is to use such accounts to illustrate the potential impact of eliminativism about conscious attitudes. Then in Section 4 we show how some of the intuitions that are naturally treated as "fixed points" in theorizing about responsibility might likewise be undermined. Section 5 draws our discussion to a conclusion, indicating some possible ways forward for theorists to consider.

We hope to establish that commitments concerning consciousness are both an important and underexplored aspect of theorizing about responsibility. We think that the general question we raise is worthy of further examination. Moreover, by emphasizing that it may be unwise for theorists to rely on the existence of conscious attitudes, we raise new issues about the purpose and relevance of many of the stock examples that have been used to motivate or constrain theories of responsibility.

1. The Importance of Consciousness

In the present section we argue that empirical discoveries about consciousness are prima facie relevant because there is, we think, ample evidence that consciousness is important to our thinking about responsibility. We should make clear that by "consciousness" throughout we mean *state* consciousness, not *creature* consciousness. It is a property that some

---

[2] We note, however, that if it should turn out that some or all libertarians are committed to the existence of conscious attitudes, then the possibility of a novel skeptical position would open up. This would be one that rejects determinism and/or endorses agent causation, but still denies that humans are responsible on the grounds that there aren't any conscious attitudes of the sort required.

[3] See S. Wolf, *Freedom within Reason* (Oxford: Oxford University Press, 1990).

mental states but not others possess (as opposed to the property that creatures who are awake and not asleep possess). In the present section we rely only on common-sense intuitions about state consciousness, making no commitment to any particular account of the latter. In Section 2 some alternative accounts will be distinguished and discussed.

The claim that responsibility requires conscious attitudes seems to be deeply entrenched in legal and moral thinking about *mens rea*, or "a guilty mind". This is because the extent to which a mind is guilty appears to be a function of the extent to which conscious attitudes are involved. Indeed, although degrees of criminal liability are now quite complex and vary somewhat between jurisdictions, it seems that our intuitive moral thinking structures the severity of a crime in proportion to the presence of conscious attitudes relevant to its commission. Premeditated crimes are often thought to be the worst, for they show conscious attention and reflection on the harm committed. Crimes committed "in the heat of the moment" aren't quite as serious, for they indicate only conscious intentionality. Finally, crimes committed out of recklessness acknowledge guilt insofar as agents are consciously aware of the risk of harm to others posed by their conduct.

Consider three parallel examples to help make the point. All involve a young man who might as well be called "Oedipus". In each case Oedipus has a conscious hatred of his father (perhaps resulting from beatings endured as a child), and in each case he does something that results in his father's death. But in one example he carefully plots his father's murder, reflecting consciously on the means to be used and the desirability of the effect. In the second example he never once entertains any conscious thoughts of killing his father, nor does he ever think to himself that it would be a good thing if his father were to die. But when an opportunity presents itself (perhaps when his father is hanging from a cliff) he acts immediately and spontaneously to cause death, without pausing to engage in any form of conscious reflection. In the third case Oedipus again never entertains any conscious thoughts involving the death of his father. But he takes a reckless decision while driving his dune-buggy at speed along a deserted beach with his father in the passenger seat, resulting in an accident that causes his father's death. Our pretheoretic intuition is that Oedipus' degree of responsibility for the outcome is significantly higher in the first of these examples than in the second, which is in turn greater than in the third. (Note that the outcome itself is the same in each case.) And the relevant differences seem to lie in the extent to which Oedipus' conscious attitudes are implicated.

It might be objected, however, that these intuitions result from the involvement of *different* attitudes, as opposed to anything to do with consciousness. Thus it might be said that in the first case, in particular, Oedipus *desires* his father's death and selects means in the service of that desire, whereas in the second case no such desire is operative. This objection requires us to believe, however, that the second Oedipus does something intentionally that he knows will result in his father's death without desiring his father's death, even though he seems to have no other purpose in mind. This is hard to accept. But even if this difficulty is set to one side, the objection doesn't appear to offer the correct diagnosis of the case. For suppose that a Freudian account of our character's action in the second of the above examples were correct. Suppose that in addition to consciously hating his father, Oedipus *also* has an unconscious desire for his father's death, grounded in jealousy of the latter's relationship with his mother. And suppose that it is this unconscious desire that prompts Oedipus to spontaneously murder his father when the opportunity presents itself. We submit that this changes our intuitions not one whit. Oedipus' degree of responsibility remains greater in the first example than in the second, although now the only difference between the cases concerns whether or not the attitudes that result in death are conscious ones.

It would appear, then, that consciousness matters to our ascriptions of responsibility and blameworthiness. Indeed, if an action can be shown to have been undertaken without consciousness, the law (and our moral practices by extension) doesn't regard it as fit for punishment. This, at any rate, is the most natural conclusion to be drawn from the case of Ben Parks, who was acquitted of murder on the grounds that he had killed his victims in a state of automatism, while sleep-walking – presumably without entertaining any conscious attitudes.[4] And in our ordinary lives, we are surely less likely to blame those who harm us via actions lacking any conscious engagement, thinking such people unfit for moral condemnation *because* their attitudes are unconscious.

This point can be dramatically illustrated through consideration of *alien hand syndrome*.[5] This is a neurological condition in which people make movements of one arm and hand that appear to be purposeful and controlled, but which the subjects themselves claim to be involuntary. Indeed,

---

[4] See R. Broughton, R. Billings, *et al.*, 'Homicidal Somnambulism: A Case Report', *Sleep* 17 (1994), pp. 253-64.

[5] See I. Biran and A. Chatterjee, 'Alien Hand Syndrome', *Archives of Neurology* 61 (2004), pp. 292-4.

those subjects will sometimes try to *prevent* the actions of the alien hand by using their other hand. (This aspect of alien hand syndrome will be familiar to many readers from its depiction in Stanley Kubrick's movie, *Dr Strangelove*.) Suppose, then, that Harry's alien hand sometimes expresses violent tendencies towards his wife. And suppose that on one occasion he isn't able to prevent the hand from acting, resulting in an injury to her. Should we hold Harry responsible? We think that intuition firmly dictates a negative answer. For the movements of his hand were outside of his conscious control. Indeed, he may have done his best to prevent the injurious action. In which case, it seems, only actions that are controlled by *conscious* attitudes are ones for which we can be held responsible.

It is important to see that this example can't be dismissed by claiming that movements of the alien hand aren't genuinely purposive. For essentially the same phenomenon sometimes occurs in cases of surgically induced division of the two brain hemispheres (commissurotomy, which involves severing the corpus callosum in the treatment of severe epilepsy). We know that in such patients movements of the left hand (under the control of the right hemisphere) are sometimes regarded by the speech-using left hemisphere as having been involuntary. But we also have every reason to believe that the right hemisphere has attitudes of its own, and initiates intentional actions on the basis of those attitudes.[6] Now admittedly, some people think that each of the two hemispheres of a commissurotomy patient realizes a numerically distinct agent, and perhaps also a distinct person.[7] But the brain damage underlying alien hand syndrome can be comparatively minor, sometimes only involving a *portion* of the corpus callosum.[8] Hence many of the activities of the two hemispheres will continue to be unified and integrated, just as in a normal person. In these cases there should be no doubt that the person in question constitutes just a single agent with a single mind, albeit an agent whose left hand sometimes acts under the control of unconscious attitudes, and a mind that contains *conflicting* attitudes.

There is some reason to think, then, that our intuitions about moral responsibility presuppose a commitment to conscious attitudes. Yet such

---

[6] See C. Marks, *Commissurotomy, Consciousness, and Unity of Mind* (Cambridge, MA: MIT Press, 1990); M. Tye, *Consciousness and Persons* (Cambridge, MA: MIT Press, 2003).

[7] R. Pucetti, 'The Case for Mental Duality: Evidence from Split-Brain Data and Other Considerations', *Behavioral and Brain Sciences* 4 (1981), pp. 93-123.

[8] See D. Geschwind, M. Iacoboni, *et al.*, 'Alien Hand Syndrome: Interhemispheric Motor Disconnection Due to a Lesion in the Midbody of the Corpus Callosum', *Neurology* 45 (1995), pp. 802-8.

commitments rarely show up explicitly in philosophical theorizing about responsibility. Indeed, few if any theories of responsibility make use of the term "conscious" in their canonical formulation. One reason for this may be that the distinction between conscious and unconscious attitudes isn't by any means at the foreground of people's thinking about the mind. Unless moral theorists have been steeped in Freudian theory, or have significant knowledge of contemporary cognitive science, then the distinction in question probably won't be salient enough to impact their thinking. In addition, if the assumption of conscious attitudes is one that crosses theoretical lines, then theorists won't feel the need to articulate it explicitly, even if they are aware of it. In Section 3, however, we will discuss "Real Self" views, showing how assumptions about consciousness can nevertheless have a powerful *in*direct impact on theorizing about responsibility. And in Sections 3 and 4 we will show how such assumptions can underwrite theorists' intuitions about key philosophical examples.

2. The Challenge to Conscious Attitudes

In the present section we embark on a brief discussion of the emerging case against the existence of conscious attitudes. This involves a wide array of theory and data from across cognitive science. Obviously, in the brief space available to us here we can't hope to convince the reader that the case in question is a strong one. Indeed, even those writers who have been most outspoken in pushing the argument against conscious attitudes concede that the considerations in question are far from conclusive at this point, and that much remains to be investigated.[9] Our goals in this section are a great deal more limited. One is to explain just enough about the theoretical perspectives involved to render the denial of conscious attitudes even so much as *intelligible*, thus putting the question on the table. Another is to indicate some of the empirical considerations that support such an account. The intended upshot is that it would be worthwhile for philosophers to consider the conditional question that forms the topic of this article, namely: If it were to turn out that there are no conscious attitudes, then what implications would this have for theories of responsibility?[10]

---

[9] See P. Carruthers, 'How We Know Our Own Minds: The Relationship Between Mind-Reading and Metacognition', *Behavioral and Brain Sciences* 32 (2009), pp. 121-182.

[10] And in any case, of course, conditional questions in philosophy can often have an interest that is independent of the plausibility of their antecedents.

2.1. Two Accounts of Conscious Attitudes

Discussions of state-consciousness in recent decades have been dominated, for the most part, by competing accounts of *phenomenal* consciousness. (This is the felt aspect, or what-it-is-like-ness, of experience.) While many authors have remained silent on the question of what makes *attitudes* conscious, this literature supports two broad possibilities. One would be modeled on first-order theories of phenomenal consciousness, of the sorts proposed by Dretske, Tye, and others.[11] It would claim that conscious attitudes are those that are "globally broadcast" or are widely available to an extensive range of systems for reasoning, decision-making, and verbal expression. The other would be modeled on higher-order theories of phenomenal consciousness, endorsed by philosophers such as Lycan,[12] Carruthers,[13] and Rosenthal.[14] It would say that conscious attitudes are those that we are *aware* of having, or that we know ourselves to possess (in the right sort of direct way). Note, moreover, that the field can be broadly divided up in this way whether or not any of the accounts in question can provide a successful *reductive* account of phenomenal consciousness. For even non-reductive theorists can accept that consciousness at least *coincides* with global broadcast (as does Chalmers,[15] for example) or with higher-order availability. And what is at stake, for our purposes, is the question whether any attitudes occupy the right sorts of causal role to count as conscious (on either of the above accounts), not whether their status as such can be reductively explained.

If we model our account of conscious attitudes on first-order theories of phenomenal consciousness, then we shall say that those attitudes are ones that are widely accessible to each other, and to processes involved in reasoning and decision-making. This comports nicely with the emphasis that many philosophers place on the "inferential promiscuity" of personal attitudes.[16] The idea is that a conscious attitude should be able to interact with

---

[11] See F. Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995); M. Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995).

[12] W. Lycan, *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996).

[13] P. Carruthers, *Phenomenal Consciousness* (Cambridge: Cambridge University Press, 2000).

[14] D. Rosenthal, *Consciousness and Mind* (Oxford: Oxford University Press, 2005).

[15] D. Chalmers, 'Availability: The Cognitive Basis of Experience', *Behavioral and Brain Sciences* 20 (1997), pp. 148-9.

[16] See G. Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982); B. Brewer, *Perception and Reason* (Oxford: Oxford University Press, 1999).

any other of one's conscious attitudes, and should be capable of being integrated with them in processes of inference of various sorts. In contrast, unconscious attitudes remain "subpersonal", and can only interact in limited ways with others, perhaps within specialized processing systems of some sort.

If we model our account of conscious attitudes on higher-order theories, in contrast, then we shall say that a conscious attitude is one that the subject is aware of having, either through the operations of some sort of a faculty of "inner sense" or via the activity of a faculty of higher-order thought. Note that most higher-order accounts of consciousness can be described as warranting a belief in *introspective* access to our own experiences (and by extension to our thoughts), in the sense that the relationship is held to be especially immediate and direct. On this view, the access that we have to our own experiences (and thoughts) is radically different from the sort of interpretative access that we have to the experiences and thoughts of other people. Indeed, this is believed by many people to be an important mark in favor of the approach. The exception to this generalization is Rosenthal.[17] He maintains that the only constraint on the way in which higher-order thoughts are generated, in order for the targeted state to count as conscious, is that it shouldn't involve any *conscious* inferences or interpretations. He thinks that provided that the interpretative process in question remain unconscious, it can take exactly the same form as the unconscious inferences that might underlie our attribution of mental states to another person.

While most people find Rosenthal's position highly counterintuitive, it should be stressed that the problem is *not* with the idea that our access to our own mental states might be *inferential*. On the contrary, inner sense theories will maintain that there can be inferences that take place within the introspective faculty, just as they take place within our perceptual faculties during external perception. But these are inferences that are supposed to be "encapsulated" from beliefs about the subject's physical circumstances, behavior, and other mental states. So a sharp divide between self-knowledge and other-knowledge is preserved by most types of higher-order theory. What is problematic about Rosenthal's account is that any such difference is erased. Not only does this require us to abandon the claim that our access to our own minds is somehow privileged and special, but one also wonders why, if interpretative access to my own mental

[17] D. Rosenthal, *Consciousness and Mind*.

states renders the latter conscious, my similar interpretative access to the mental states of other people shouldn't also render *them* conscious.[18] In the discussion that follows, therefore, we assume that if propositional attitudes are to count as conscious, according to a higher-order account, then our access to those attitudes must be non-interpretative, and must occur independently of any beliefs about our own circumstances, behavior, or other mental states.

It appears, therefore, that if there exist any conscious attitudes, then they would either be globally broadcast and promiscuously available to other such attitudes and decision-making processes, or they would be ones of which people have non-interpretative awareness. There is, however, a third possibility represented in the literature on phenomenal consciousness. This is biological as opposed to functional in character, and is most famously defended by Block.[19] On this sort of account, phenomenal consciousness is identical with (or in a non-reductive formulation, coincides with) some unknown set of neural properties of specific areas of the brain. It is therefore left open that there can be phenomenally conscious states that are inaccessible to their subjects in both a first-order and a higher-order sense. An account of conscious attitudes modeled on this approach would claim that such attitudes can fail to be globally broadcast, as well as claiming that they can be ones that subjects aren't aware of having. While such a view is conceptually coherent, it is hard to see anything that might motivate it. Nor has anyone yet attempted to defend it. We therefore set this possibility to one side in the discussion that follows.

2.2. *Attitudes and Global Broadcast*

Here we suggest that there is accumulating evidence from cognitive science that attitudes, as such, aren't globally broadcast or promiscuously

[18] This is a variant of the so-called "rock objection" to higher-order theories: if awareness of a mental state renders it conscious, then how is it that awareness of a rock doesn't render it conscious? See A. Goldman, 'Consciousness, Folk-Psychology, and Cognitive Science', *Consciousness and Cognition* 2 (1993), pp. 364-82; L. Stubenberg, *Consciousness and Qualia* (Amsterdam: John Benjamins, 1998). But the present argument isn't vulnerable to the obvious rejoinder, which is that only mental states are the right *kinds* of thing to be conscious. See W. Lycan, *Consciousness and Experience*.

[19] N. Block, 'A Confusion about the Function of Consciousness', *Behavioral and Brain Sciences* 18 (1995), pp. 227-47; N. Block, 'The Harder Problem of Consciousness', *The Journal of Philosophy* 99 (2002), pp. 1-35.

available. Hence if one's preferred account of conscious attitudes is first-order, then it is beginning to look as if there might be no such things.

An initial point is that all of the evidence that we have of global broadcasting in the brain concerns perceptual events, together with events that use the very same mechanisms as perception, such as visual and auditory imagery.[20] It might be replied, however, that absence of evidence isn't the same thing as evidence of absence. Hence it remains possible that there is a centralized working-memory workspace within which propositional attitudes can be activated and engaged with one another, and the contents of which are made globally accessible to all or most of the concept-wielding executive systems in the mind for belief-formation, reasoning, and decision making. However, the best-established model of working memory permits no such thing. This is the theory developed and experimentally investigated over the years by Baddeley and colleagues.[21] On this account, the working memory system consists of a central executive which directs and utilizes two "slave" systems – the phonological loop and the visuo-spatial sketchpad – together with an "episodic buffer" that serves to integrate the two sensory slave systems with information from semantic and episodic memory, binding them together.[22] Crucially, there is no suggestion that the central executive of the system can function in the *absence* of the slave subsystems. So although the working memory system is, indeed, a kind of

---

[20] See, e.g., S. Dehaene and L. Naccache, 'Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework', *Cognition* 79 (2001), pp. 1-37; B. Baars, 'The Conscious Access Hypothesis: Origins and Recent Evidence', *Trends in Cognitive Sciences* 6 (2002), pp. 47-52; B. Baars, T. Ramsoy, *et al.*, 'Brain, Consciousness, and the Observing Self', *Trends in Neurosciences* 26 (2003), pp. 671-5; S. Dehaene, J-P. Changeux, *et al.*, 'Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy', *Trends in Cognitive Sciences* 10 (2006), pp. 204-11.

[21] See, e.g., A. Baddeley and G. Hitch, 'Working Memory', in G. Bower (ed.), *Recent Advances in Learning and Motivation*, vol. 8 (New York: Academic Press, 1974); A. Baddeley and R. Logie, 'Working Memory: The Multiple-Component Model' in A. Miyake and P. Shah (eds.), *Models of Working Memory* (Cambridge: Cambridge University Press, 1999); A. Baddeley, *Working Memory, Thought, and Action* (Oxford: Oxford University Press, 2006).

[22] The phonological loop activates and maintains linguistic representations, or so-called "inner speech". The visuo-spatial sketchpad is responsible for broadcasting visual images. In light of the recent discovery of the important role played by motor imagery in conscious learning and reasoning (see M. Jeannerod, *Motor Cognition* (Oxford: Oxford University Press, 2006)), a third slave system should probably be added. (Indeed, see P. Barnard, 'Interacting Cognitive Subsystems' in A. Miyake and P. Shah (eds.), *Models of Working Memory*, for just such a proposal.)

global workspace, it isn't one in which *attitudes* interact with one another. (Or at least, they only do so via their formulation into visual imagery or inner speech. We return to this point shortly.)

Moreover, if attitudes were globally accessible to all conceptual systems for belief formation, decision making, and so forth, then one would expect them also to be available as input to the common-sense psychology faculty (often now described as the "mindreading" system). In which case our activated attitudes should be capable of self-attribution immediately and without interpretation. But as we will see in Section 2.3, the evidence suggests that this is not the case. If this turns out to be so, then that will render it highly unlikely that attitudes are globally broadcast. For the suggestion that they might be broadcast to all conceptual systems with the exception of the mindreading faculty is at best *ad hoc* and at worst bizarre. For on almost everyone's view, higher-order (or "metacognitive") thoughts are supposed to play an important role within the central-process (or "executive") operations of the human mind.

What, then, are we to make of the common-sense data that leads philosophers to believe in the "inferential promiscuity" of conscious attitudes, or to believe in a centralized workspace in which all of one's attitudes can interact with and influence one another? Isn't it true that any belief of ours can, in principle, be brought to bear in the evaluation of any other, as Fodor[23] argues? Indeed, there is a sense in which it *is* true. But the evidence suggests that the manner in which any one of our attitudes can gain global access to the remainder is indirect, and is dependent upon it being used to generate visual imagery, on the one hand, or on formulation into inner speech, on the other. It is the resulting imagistic representations that are globally broadcast, not the attitudes themselves.

2.3. Higher-Order Knowledge of Attitudes

Here we suggest that there is a growing body of evidence in cognitive science for the conclusion that our only access to our own attitudes is interpretative in character, much like our access to the attitudes of other people. Hence if conscious attitudes are supposed to be those that we have non-interpretative higher-order access to, then it might well turn out that there aren't any.

---

[23] J. Fodor, *The Mind Doesn't Work That Way*, (Cambridge, MA: MIT Press, 2000).

Such claims are supported by the following theoretical perspective.[24] There is just a single faculty of the human mind underlying our capacity to attribute mental states, whether to others or to ourselves. This faculty initially evolved (or the mechanism to acquire it via learning evolved) for purposes of social interaction, enabling us to anticipate and manipulate the behavior of other people, as well as to cooperate more effectively with them.[25] In order to do its work of interpreting the behavior of others, the mindreading faculty needs to have access to the outputs of perception (thereby also gaining access to various forms of imagery, which share the same mechanisms). Indeed, the mindreading system is located as one of a suite of conceptual systems for generating emotions, motivations, judgments of risk, judgments of physical causality, and so on, which are all arranged as consumers of the globally broadcast outputs of the various

---

[24] For further details, see Carruthers, 'How We Know Our Own Minds,' and P. Carruthers, *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (Oxford: Oxford University Press, 2011).

[25] One approach stresses manipulation. This is the so-called "Machiavellian intelligence hypothesis." See R. Byrne and A. Whiten (eds.), *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans* (Oxford: Oxford University Press, 1988) and R. Byrne and A. Whiten (eds.), *Machiavellian Intelligence II: Extensions and Evaluations* (Cambridge: Cambridge University Press, 1997). Another stresses cooperation. See M. Tomasello, *Origins of Human Communication* (Cambridge, MA: MIT Press, 2008) and S. Hrdy, *Mothers and Others* (Cambridge, MA: Harvard University Press, 2009).

Consistent with such claims, there is now significant evidence of primitive mindreading abilities in other highly social creatures, especially monkeys and apes. See B. Hare, J. Call, *et al.*, 'Do Chimpanzees Know What Conspecifics Know?' *Animal Behavior* 61 (2001), pp. 139-51; B. Hare, E. Addessi, *et al.*, 'Do Capuchin Monkeys, *Cebus paella*, Know What Conspecifics Do and Do Not See?' *Animals Behavior* 65(2003), pp. 131-42; B. Hare, J. Call, *et al.*, 'Chimpanzees Deceive a Human Competitor by Hiding', *Cognition* 101 (2006), pp. 495-514; M. Tomasello, J. Call, *et al.*, 'Chimpanzees Understand Psychological States – the Question is Which Ones and to What Extent', *Trends in Cognitive Sciences* 7 (2003), pp. 153-6; J. Flombaum and L. Santos, 'Rhesus Monkeys Attribute Perceptions to Others,' *Current Biology* 15 (2005), pp. 447-52; L. Santos, A. Nissen, *et al.*, 'Rhesus Monkeys (Macaca mulatta) Know What Others Can and Cannot Hear', *Animal Behavior* 71 (2006), pp. 1175-81.

Likewise there is increasing evidence that mindreading capacities are innately channeled in human infants, emerging early and reliably in the first year or two of life. See G. Csibra, S. Bíró, *et al.*, 'One-Year-Old Infants Use Teleological Representations of Actions Productively', *Cognitive Science* 27 (2003), pp. 111-33; K. Onishi and R. Baillargeon, 'Do 15-month -Olds Understand False Beliefs?' *Science* 5719 (2005), pp. 255-8; V. Southgate, A. Senju, *et al.*, 'Action Anticipation Through Attribution of False Belief by 2-Year-Olds', *Psychological Science* 18 (2007), pp. 587-92; H. Song and R. Baillargeon, 'Infants' Reasoning About Others' False Perceptions', *Developmental Psychology* 44 (2008), pp. 1789-95; D. Buttelmann, M. Carpenter, *et al.*, 'Eighteen-Month-Old Infants Show False Belief Understanding in an Active Helping Paradigm', *Cognition* 112 (2009), pp. 337-42.

perceptual systems.[26] As a result, attributing perceptual and imagistic events to ourselves is easy, and certainly doesn't need to involve interpretations of our own behavior. Indeed, the model confirms our intuition that we have immediate, introspective, access to our own visual and auditory perceptions, as well as to our own visual images, "inner speech", and so forth.

However, on this account the mindreading system has *no* access to the subject's own propositional attitude events of judging, deciding, endorsing, and so on. These can only be self-attributed by turning the mindreading system's interpretative powers onto the self. Notice, however, that the data available for interpretation will include not just perceptions of the agent's own actions and circumstances, but also the subject's own visual images, inner speech, emotional feelings, and so forth, which are presented as input to the mindreading system via global broadcast. Naturally, this will result in a significant increase in the reliability of self-attribution as against other-attribution. But since our own attitudes would still only be known via self-interpretation they wouldn't qualify as conscious, according a higher-order account.

The theory just sketched predicts that people should be led to *confabulate* (i.e. falsely attribute) attitudes to themselves whenever they are presented with misleading behavioral, circumstantial, or sensory evidence (just as they can be caused to misinterpret other people in such cases). There is now a wealth of data from cognitive science using many different paradigms that this is, indeed, the case.[27] Here we will sketch just one strand of evidence for purposes of illustration.

It has long been known that subjects who are induced to nod their heads while listening to a tape via headphones (ostensibly to test the headphones themselves) will say that they have a greater degree of belief in the propositions being defended on the tape than will subjects who are induced to shake their heads.[28] It seems that subjects reason: "Since I am nodding / shaking my head, this is evidence that I believe / disbelieve the propositions asserted." Admittedly, this isn't the only possible explanation. It

---

[26] See A. Baars, *A Cognitive Theory of Consciousness* (Cambridge: Cambridge University Press, 1988); M. Shanahan and A. Baars, 'Applying Global Workspace Theory to the Frame Problem', *Cognition* 98 (2005), pp. 157-76; P. Carruthers, *The Architecture of the Mind* (Oxford: Oxford University Press, 2006).

[27] For extensive review and discussion, see P. Carruthers, 'Introspection: Divided and Partly Eliminated', *Philosophy and Phenomenological Research* 80 (2010), pp. 76-111.

[28] G. Wells and R. Petty, 'The Effects of Overt Head Movements on Persuasion', *Basic and Applied Social Psychology* 1 (1980), pp. 219-30.

might be that head-nodding primes for positive thoughts about the message, which in turn cause greater agreement, which is then introspected and veridically reported. Briñol and Petty[29] set out to test this alternative by varying the persuasiveness of the messages themselves. When the message is persuasive, nodding increases belief and head-shaking decreases it, which is consistent with either one of the two explanations just sketched. But when the message is *un*persuasive the opposite occurs: nodding *decreases* belief and head-shaking *increases* it. This isn't consistent with a priming-for-positive-thoughts account. The authors present evidence that what is actually happening is that subjects interpret their nodding behavior as confirming their own initial negative reactions to the message, while head-shaking is interpreted as disagreement with those reactions.

From the fact that we sometimes engage in swift unconscious self-interpretation it doesn't follow that we always do, of course. On other occasions we might have access to our attitudes that is introspective in character. It is important to see, however, that from the subject's own perspective there is no epistemic difference between the two. Since it can seem to subjects that they are introspecting their attitudes when they are demonstrably (but unconsciously) self-interpreting, the common-sense intuition that we have immediate introspective access to our own propositional attitude events of deciding, judging, and so forth should arguably be given little weight. Moreover if this sort of "dual method" theory of self-knowledge is to be defensible, then some account needs to be given of the circumstances in which each is used. But arguably the patterning found in the data on confabulation cannot be explained by a dual method theorist in anything other than an *ad hoc* way.[30]

One of us reviews such competing accounts of our knowledge of our own attitudes at length, comparing their predictions concerning the course of cognitive evolution, childhood development, metacognitive processes in adults, the reliability of self-report, expected dissociations resulting from targeted brain damage, and more.[31] The tentative conclusion drawn is that the evidence best supports a solely self-interpretative account of our access to our own propositional attitudes. If that conclusion turns out to be correct, then it will follow that there are no conscious attitudes. And this will be so, whether one endorses a first-order or higher-order account of attitude consciousness. For as we noted in Section 2.2, first-order accounts,

---

[29] P. Briñol and R. Petty, 'Overt Head Movementls and Persuasion: A Self-Validation Analysis', *Journal of Personality and Social Psychology* 84 (2003), pp. 1123-39.

[30] Carruthers, 'Introspection: Divided and Partly Eliminated'.

[31] Carruthers, 'How We Know Our Own Minds', and Carruthers, *The Opacity of Mind*.

too, predict that we should have non-interpretative access to our own conscious attitudes.

We close this section by reiterating the point we made at the outset. Although we have sketched some reasons for denying the existence of conscious attitudes, resulting from a variety of theories and kinds of evidence in cognitive science, we don't expect the reader to be convinced. That has not been our purpose. We hope only to have shown that the antecedent of our main conditional question makes good enough sense for the question itself to be worth considering.

3. Real Selves, Conscious Selves

In the present section we show that so-called "Real Self" theories of moral responsibility[32] are likely to be undermined if it should turn out that there are no conscious attitudes. This will serve to illustrate our general point about the potential vulnerability of theories of responsibility. We want to emphasize, however, that although Frankfurt and Watson happen to be compatibilists, the issues that they raise and the examples that they use should be equally accessible from many libertarian perspectives. In fact the concern that lies at the heart of Real Self views would seem to be one that cuts across the traditional dispute between compatibilists and libertarians. It is the question of how to identify within the subject the proper locus of responsibility.

3.1. Cartesian Epistemology and the Real Self

We open this discussion with a parable. There was once a philosopher named "John", who began his career as a naïve Cartesian about mental event epistemology. (We will assume that he was neutral on questions of ontology, or else that he was a convinced physicalist; this isn't relevant to what follows.) John initially believed that minds are transparently accessible to themselves. He might even have been prepared to assert, as Locke[33]

[32] For canonical versions, see H. Frankfurt, 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* 68 (1971), pp. 5-20 and G. Watson, 'Free Agency', *Journal of Philosophy* 72 (1975), pp. 205-20. For further developments, see H. Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988) and G. Watson, *Agency and Answerability* (Oxford: Oxford University Press, 2004).

[33] J. Locke, *An Essay Concerning Human Understanding* (many editions, 1690).

once did, "There can be nothing within the mind that the mind itself is unaware of", except that John was alive to the distinction between dormant mental states (standing beliefs, stored memories, long-term motives, and so forth) and active events (judging, experiencing, remembering, feeling, and so on). For it is part of common sense that the former can sometimes be hard for us to access. We can know that we know something, for example, although right now we can't bring it to mind. What John believed is that all of his own mental *events* are transparently accessible to him. He couldn't think something, or judge something, or remember something, or feel something without awareness that he was doing so; and his beliefs about his own mental events weren't grounded in any sort of *interpretation* (as were his beliefs about the mental events of other people) but were rather *immediate* (albeit not infallible, perhaps – John wasn't a true Cartesian).

Among his mental events, John distinguished between those that simply *arrived* in his mind from external causes (like feelings of pain or hunger) and those that were the product of his own mental activity (like decisions produced by his own reasoning and reflecting). For the latter he was inclined to claim responsibility, of course. Externally caused events, in contrast, were seen (initially, at least) as impositions upon him from outside. But once endorsed and/or sustained by reflection he was inclined to think of them, too, as genuinely his own, and to take responsibility for them. Indeed, John came to think that *actual* endorsement wasn't necessary for a mental state to be genuinely his own. It was enough that a desire or thought should be one that he *would* endorse on reflection if the question were ever raised.[34]

Then John discovered cognitive science (or Freudian psychology, or both). He came to believe that there are unconscious perceptions, judgments, decisions, and emotions. And he learned that there are all sorts of unconscious cognitive processes underlying the genesis of the events that take place in his conscious mind. How did he respond? Under pressure from his pre-theoretic Cartesianism, he was inclined to think that only the mental events of which he was aware could genuinely be *his*, whereas the mental events that cognitive science talks about were within his *body*, perhaps, but were not really part of his *self*. (Perhaps he introduced the

[34] In what follows we use "endorse" as a semi-technical term, intended to be neutral between the Real Self accounts of Frankfurt and Watson. Roughly, one can endorse a desire *either* by having a desire that the desire in question should motivate one's actions *or* by judging that the desire in question is consistent with one's values.

language of "personal" versus "sub-personal" mental states to mark the distinction.)[35] And when he turned his mind to questions of responsibility, he was inclined to think that conscious events whose causes were perhaps mental in character, but unconscious, were imposed upon him from outside (just like feelings of pain and hunger). Only when he was aware of the genesis of an event through a process of conscious reflection did he feel that *he* had caused the resulting state. And only when he was inclined to give conscious endorsement to events that weren't so caused did he feel that they were appropriated into his self, and did he wish to take responsibility for them or their effects.

The point of the parable is that John stands for the folk, and especially for the line of thought that gives rise to Real Self accounts of responsibility, as we will demonstrate below. One of us has argued that the Cartesian conception of minds as transparently accessible to themselves (which is where John started out) might be an innately channeled aspect of human social cognition.[36] Certainly that conception has a powerful pre-theoretical appeal. Moreover, recall that on either of the two approaches to attitude consciousness distinguished in Section 2, all and only conscious attitudes should be non-interpretatively accessible to their subjects, just as a Cartesian account predicts. Hence if one reasons as John did, then it will be extremely natural to think that only conscious mental events that are caused by conscious attitudes or are otherwise consciously appropriated (or are such that they *would* be consciously endorsed if reflected upon) are genuinely our own, and can serve as a locus of responsibility.

3.2. Compulsion and Our "Real" Selves

Here we lay out some of the examples and considerations that have actually been used to motivate Real Self views of responsibility, initially without using the term "conscious". We then argue that the theorists in question must be tacitly assuming that the attitudes appealed to in such accounts are conscious ones.

A powerful intuitive case in support of Real Self views concerns victims of compulsion. Consider Mr Klepto, a kleptomaniac, who often feels a

---

[35] See J. Hornsby, *Simple Mindedness* (Cambridge, MA: Harvard University Press, 1997); J. Bermúdez, 'Personal and Subpersonal', *Philosophical Explorations* 2 (2000), pp. 63-82.

[36] P. Carruthers, 'Cartesian Epistemology: Is the Theory of the Self-Transparent Mind Innate?' *Journal of Consciousness Studies* 15:4 (2008), pp. 28-53; Carruthers, *The Opacity of Mind*.

powerful urge to steal. Klepto's desire presents itself to him as an unwanted urge, a force that he must fight to resist. Should he succumb, and steal something as a result, it appears that he does so unfreely, and is not responsible. Indeed, one might be tempted to claim that accounting for Klepto's non-responsibility is a fixed data-point for any theory of responsibility to capture.

One natural way to capture both the phenomenology of Klepto's situation and to show why he is not responsible is to equate free and responsible action with those actions produced by attitudes that belong to an agent's "real" or "genuine" self.[37] Some actions reveal our genuine values, commitments, and motivations, while others are performed in opposition to our genuinely held values, commitments, and motivations. According to this view, it is only the former that are free and for which we are responsible. Our real selves are determined not by what we desire, but either by the motivations that we want to motivate us or by our considered value judgments. To be responsible for an action requires that the action be motivated by a desire that has been (or would be) *endorsed*, for it is such endorsements that disclose our real selves. Thus, a smoker who has judged continued smoking to be bad for his health and has, as a result, committed to quitting (thereby valuing his health above temporary pleasure), and who nevertheless acts on his urge to smoke another cigarette, does not act freely. The agent's real self is here a harmonious self; only when the agent acts on a desire that is (or would be) endorsed does he act freely.

Real Self views capture the experience of inner struggle in a satisfying way by showing how the real self is the locus of agency and the source of free and responsible action. Compulsive desires and temptations against our better judgments are obstacles for our self to overcome, not its proper products. This is why Frankfurt[38] appeals to the unwilling addict to illustrate the importance of reflective endorsement. The unwilling addict can't help desiring to take the drug. But he is an *unwilling* addict because he has rejected that desire. He doesn't want to take the drug, but more importantly he doesn't want his desire to take it to motivate him. If he ends up succumbing to his desire to take the drug and so takes it, he will be acting against his judgment not to be motivated by his desire to take it. Since it is one's reflective judgments (actual or potential) that characterize one's real self, the unwilling addict doesn't freely take the drug and isn't responsible for doing so.

---

[37] Frankfurt, 'Free Will and the Concept of a Person'; Watson, 'Free Agency'.

[38] Frankfurt, 'Free Will and the Concept of a Person'.

In contrast, the *willing* addict has the same desire to take the drug, but he differs in that he has endorsed having that desire and thus appropriated it. Hence on this view, when he takes the drug he does so freely and is responsible for doing so. Here the initial origin of the willing addict's desire for the drug must be the same as the unwilling addict's. What makes the cases different is not the origin of the motivation, but its relation to the agent's real self: in the case of the unwilling addict, the desire for the drug is alien to his real self; in the case of the willing addict, that desire has been appropriated by his real self.

While some mental states therefore belong to the real self through appropriation, there are other states that belong by *originating within* the real self. These are the states one gets to through reflective reasoning, being caused by that process. While endorsement appropriates desires (whose origins lie elsewhere), the states that constitute reflective endorsement are themselves the *products* of the real self. (Instances of *un*reflective endorsement, in contrast, can belong to the real self provided that they would, in turn, be reflectively endorsed.) Reflective endorsements belong to the real self *by* being the products of reflective reasoning. It follows then that the real self is to be identified with the locus of processes of reflective reasoning.

It is worth noting, however, that despite talk of attitudes belonging to the real self or being produced by it, the real self view is not to be understood as making any ontological commitments. All that need be meant by reference to a real "self" is a collection of states and attitudes and the processes that govern or produce them. The notion of a real self is intended just to mark the distinction between those attitudes that are relevant to the moral evaluation of individuals and those that are not.

In our view, if the real self is to serve the role of distinguishing between mental states that are alien and those that are genuinely one's own by either appropriation or genesis, then it must contain *conscious* attitudes. Certainly the parallel between Real Self views and the parable of John from Section 3.1 is striking. We submit that although neither Frankfurt nor Watson uses the language of consciousness, the reflections, higher-order motives, and value judgments that characterize an agent's real self (on such views) must be conscious ones.

Consider Mr Klepto again. He often experiences an intense urge to steal items from the stores that he visits. On the proposed view, he steals freely and is responsible only if his desire to steal belongs to his real self; that is, only if he wishes that his desire to steal be operative, or if stealing is something that he values. The intuitive appeal of this requirement is that only

his endorsements accurately reflect his commitments as an agent. His impulses to steal are external impingements on his freedom. If he reluctantly succumbs to them, then he doesn't act freely, for his will is bent to the mercy of strong "alien" desires. Only when he willingly adopts these impulses, or when they accord with his values, should we take his act of stealing to be free and responsible; only then should we take it to express his real self. While these claims don't explicitly require that such endorsements should be conscious, it is clear that they must be if we are to retain the intuitive appeal of the view.

To see this, suppose that Klepto's urge to steal is actually appropriated by an *unconscious* endorsement. And suppose that he acts on that desire without further reflection and so isn't conflicted. Surely this sort of endorsement would be insufficient to secure Klepto's autonomy and responsibility for stealing, especially if conscious reflection on his thieving desires would have had no impact. If Klepto steals based on a desire that he unconsciously wishes to be operative, or in accordance with an unconscious evaluation, then there is no reason to suppose this is any more indicative of his real self than if he had succumbed to the initial desire unwillingly.

Alternatively, suppose that some or other variant of Freudian psychology is correct, and that certain desires (e.g., to get back at my father) are generated *and* endorsed within the unconscious mind. If I act on such a desire then I act on an endorsed desire, but it is implausible to suppose that it reflects my real self. Such endorsement clearly doesn't capture the intuitive appeal of the examples, which motivate equating the responsible self with a self containing conscious attitudes. For it would seem that desires receiving unconscious endorsement would be equally "outside" the real self as pathological compulsions, and an appropriation system or evaluation system that bypassed conscious thought would by all appearances bypass the real self. Indeed, from Klepto's perspective, unconscious endorsement would seem just as alien to him as the initial urge itself.[39]

It seems, therefore, that no desire could count as belonging to one's real self unless *consciously* desired to be operative, or cohering with *conscious* value judgments. Unconscious processes bear greater resemblance to

---

[39] Compare Nahmias, who stresses that to the degree that we are influenced by unconscious factors (of which we are not aware) we are less responsible for what we do as a result. His argument, too, suggests that conscious attitudes are important to responsibility. See E. Nahmias, 'Autonomous Agency and Social Psychology', in M. Marraffa, M. Caro, and F. Ferretti (eds.), *Cartographies of the Mind: Philosophy and Psychology in Intersection* (Dordrecht: Springer, 2007).

"outside" and "alien" forces than to processes proper to the real self. Moreover, if there were a struggle between consciously endorsed desires and unconsciously endorsed ones, we would of course identify ourselves, our *real* selves, with the former set. So the mere fact that a desire receives support from one's other desires or value judgments cannot suffice for belonging to the real self. The endorsement in question needs to be conscious.

3.3. The Threat to the Real Self

If the model of our psychology sketched in Section 2 proves to be on the right lines, and if we are correct in our surmise that the states that constitute a real self must be conscious ones, then it will follow that there is no such thing as a real self. For a self, to count as such, presumably needs to contain a full range of mental-state types. It should contain, not only perceptual states, but also attitude states. But it would have turned out that there are no such things as conscious events of judging, deciding, endorsing, and so on. There are, of course, judgments, decisions, and endorsements. But these always occur unconsciously, below the surface of awareness. We only ever know that they are occurring by interpreting our own overt and covert behavior (including so-called "inner speech"), together with our own perceptions, emotional feelings, and so forth.

This isn't to say that our propositional attitudes never have conscious *effects*, of course. And included among the latter might be their expression in inner or outer speech. This means that some of the events that occur in us consciously will take the superficial form of propositional attitudes, and might naturally be (mis)identified as such. One might find oneself saying in inner speech, for example, "It is good that I like salads", or "I wish that I could give up smoking." Awareness of events such as these makes it especially natural for us to think that we have non-interpretative awareness of our own acts of reflective endorsement, among other attitudes. But although natural, such a belief may be illusory. Even if one's speech acts (whether covert or overt) are more or less reliably *caused* by one's attitudes, the introspective access that one has to the effects of the former (inner or outer speech) doesn't give one introspective access to the latter. On the contrary, inner speech, just like any other form of speech, needs to be *interpreted* for one to have access to the attitudes that caused it.[40]

---

[40] Carruthers, 'How We Know Our Own Minds'; Carruthers, *The Opacity of Mind*.

Moreover speech, like any other form of action, can be undertaken for a range of purposes in addition to the expression of one's attitudes. And included among the purposes for which we use inner speech are self-exhortation ("I shouldn't have another cigarette") and the eliciting of information through self-questioning ("What would be good to eat?"). Indeed, many theorists have claimed that a whole new system for thinking and reasoning can be built out of sequences of inner verbalization, guided by learned habits, acquired beliefs about how one should reason, and so forth.[41] And sometimes, as a result, a rehearsed sentence in inner speech can have a causal role *a bit like* that of a judgment, or a decision. Thus it might be because one says to oneself, "I shouldn't have a cigarette" that one declines the offer of a cigarette, somewhat as if the verbalization were to constitute a decision not to have one. But still the conscious event in question doesn't have the right kind of causal role to *be* a decision.

Consider how inner verbalization achieves its characteristic effects in cases like this. According to Frankish[42] it happens via us interpreting ourselves as having undertaken various sorts of *commitment*. Hearing oneself rehearse the words, "I shouldn't have a cigarette", for example, one (that is to say: one's mindreading system) might interpret oneself as having made a commitment not to have a cigarette. Then this belief, together with a standing desire to execute one's commitments, provides a motive for rejecting the proffered cigarette, which might on this occasion be sufficient to issue in a decision to do so. In this scenario, of course, it isn't the saying of the words, "I shouldn't …" that constitutes one's decision. Rather, the real decision is taken down-stream of that event, once one has interpreted oneself as having made a commitment, in the presence of a desire to execute commitments.[43]

3.4. Klepto et al. Revisited

We now consider how the cases that motivate the idea of a real self would appear from this revised perspective. Take a conflict case: one becomes

---

[41] See, e.g., D. Dennett, *Consciousness Explained* (Boston: Little, Brown and Co., 1991); K. Frankish, *Mind and Supermind*, (Cambridge: Cambridge University Press, 2004).

[42] Frankish, *Mind and Supermind*.

[43] In other cases, of course, the real decision may be taken up-stream, rather than down-stream, of an event of conscious inner verbalization. But the latter will only give us interpretative access to the former, and hence the decision itself won't be a conscious one.

aware of a craving to smoke a cigarette. This activates (unconsciously) various items of knowledge, such as that smoking is bad for one's health, as well as various goals (such as a desire to remain healthy). As a result of some unconscious cognitive processing one utters, in inner speech, the sentence, "I wish I could stop smoking." This is interpreted (unconsciously) as a commitment to the goal of giving up smoking, and interacts with one's standing (unconscious) desire to execute one's commitments to produce an (unconscious) desire to do things that might result in stopping smoking. In consequence one might apply a nicotine patch, or one might do nothing, but feel bad about it when one smokes the next cigarette. Here there is a conflict alright, but it is a conflict between an initial craving and a desire that remains unconscious. The first is activated by the sight of a cigarette packet (say), while the second is generated by an *action* of the subject (such as saying to oneself, "I ought to give up"). There is no reason why the latter should be considered as belonging to one's "real" self any more than does the former. For Klepto's occurrent desire to steal, too, will be caused by an action of his, such as entering the shop, or visualizing himself entering the shop. Why should desires that happen to be caused by our own actions be counted as belonging to our real selves?

Likewise in the case of an unconflicted smoker: he, too, feels the urge to smoke. At that point he might recall his mother urging him to give up, but (as a result of further unconscious processing) he verbalizes to himself, "It is my business what I do with my life; I like smoking; I want to go on." This is (unconsciously) interpreted as a commitment to the goal of continuing smoking, which interacts with his standing desire to execute his commitments to produce an (unconscious) desire to continue. So in this case the agent ends up with an unconscious motivation in support of his initial urge, but one that is caused by an action of his. Yet why should this make any difference to its status? If Klepto goes to the shop knowing that this will cause an urge to steal, but also because he wants a new watch, then he, too, ends up with two distinct motivations in support of taking the watch, one of which (the urge to steal) is caused by an action of his. But this wouldn't suddenly turn him into an autonomous agent, from the perspective of a Real Self account.

The real self (indeed, the only self that there is) would actually be an unconscious self, on the account sketched here. While there are many conscious mental *events*, of course, these are all perceptual or quasi-perceptual in character, including conscious percepts, visual images, auditory experiences, "inner speech", felt urges, and so forth. But these events wouldn't

constitute a self, since they fail to include propositional attitude events of judging, deciding, endorsing, and so on. The only propositional attitude events that would exist are *un*conscious ones, albeit many of them either causing or caused by conscious perceptual or quasi-perceptual episodes of one sort or another. There would therefore seem to be no basis on which to distinguish between events that belong to one's real self and those that don't. All of one's mental events are equally one's own, although they have a heterogeneity of causes (bodily states, external events, unconscious cognitive processing, conscious perceptions, conscious episodes of inner speech, and so on).

From this perspective, then, there would seem to be no good grounds for denying that Klepto is responsible for his actions. For his actions proceed just as much from *him*, from his one-and-only self, as do the actions of any other person. We admit that this is counter-intuitive. But if some of the views outlined in Section 2 turn out to be correct, then the intuition in question will be very much up for grabs. We will elaborate on this point in Section 4.

We should briefly note one possible response to the views sketched here. A Real Self theorist might simply deny that endorsements must be conscious, thereby giving up any commitment to conscious attitudes. But if our arguments in the present section have been sound, then this would at least be a significant revision to Real Self views as they are typically conceived. Moreover, and more importantly, any such revised Real Self account would abandon its prime motivation, which is to capture the phenomenology of cases like Mr Klepto, alien hand syndrome, and the conflicted smoker. The result is a dilemma: either Real Self accounts are committed to conscious attitudes, or the sorts of cases that best motivate them no longer provide them with support. And if Real Self views face such problems, then that suggests that other views, too, should be examined to see how they would fare in face of the discovery that there are no conscious attitudes.

As a brief but suggestive note, we add that *many* views aim to give an account of the sort of control over one's actions necessary for us to be responsible for them. In light of our present discussion, it seems natural to ask whether such control must be conscious. Admittedly, it isn't obvious that conscious *control* must require the existence of conscious *attitudes*. But neither is it obvious that it doesn't. Indeed, it seems plausible that one doesn't consciously control one's choice of the red shoes over the blue, if that decision is unconscious and only later consciously interpreted in inner speech.

4. Intuitions Re-examined

It appears that if the argument sketched in Section 2 for denying the existence of conscious attitudes were to be sustained, then the intuition that those acting from compulsive desires aren't truly responsible would be threatened. And it appears, likewise, that the intuitions articulated in Section 1 would have to be given up. For then the difference between the case where Oedipus plots his father's death and the one where he commits the killing on the spur of the moment will just be a difference in the mechanisms through which Oedipus' unconscious attitudes achieve their effects (in the one case proceeding via instances of inner speech, say, and in the other case not). And it is quite unclear why *this* difference should have any moral significance. Similarly, the fact that the motives that drive Harry's alien hand to strike his wife are unconscious ones would no longer be a reason for discounting their significance. For *all* attitudes would be similarly unconscious, differing only in whether or not they cause (or are caused by) conscious events (such as globally broadcast visual images or inner speech), in addition to issuing in bodily action. In the present section we consider whether these consequences wreak too much havoc on our intuitions to be acceptable.

Philosophers have traditionally thought that intuitions provide a reliable guide to the truth (at least when suitably isolated and clarified, as in the Cartesian doctrine of "clear and distinct perception"). Some philosophers (of a Platonic bent) have thought that intuitions are forms of intellectual insight into the structure of reality. Others have thought that they reflect the structure of our conceptual system. And yet others have thought that they are entailments of our system of beliefs. From any of these perspectives, intuitions can reasonably be taken as (almost) "fixed points" in philosophical theorizing, needing at least to be brought into reflective equilibrium with explanatory theories. They are the primary data that philosophical theories need to accommodate and explain.

This traditional picture has come under increasing pressure over the last decade. Many philosophers have become suspicious of the sources of intuition. As products of human cognition, their genesis is likely to be as much subject to bias and error as any other form of belief. Some have stressed that the apparent cultural variability of intuitions casts doubt on their central role in philosophical theorizing.[44] Others have begun to conduct

[44] E. Machery, R. Mallon, *et al.*, 'Semantics Cross-Cultural Style', *Cognition* 92 (2004), pp. 1-12.

experiments to see how widespread among people some important intuitions actually are.[45] And some have thought to ground intuitions in features of our cognitive systems that plainly have nothing to do with rational insight, conceptual structure, or even belief. Thus Scholl[46] shows how some of the key intuitions in philosophical debates about object-identity – particularly the central importance of spatiotemporal continuity and persistence through change – may actually be produced by processing structures within the early visual system. For the visual system appears to make just such assumptions independently of our beliefs and concepts (indeed, in a way that is encapsulated from the latter), thus issuing in powerful intuitions whenever we visually imagine any of the standard examples involved in debates over object identity.

We suggest that some of the intuitions involved in debates over responsibility may likewise reflect the processing structures of one particular cognitive faculty; in this case the mindreading system. If that system operates with the tacit assumption that minds are transparent to themselves, as Carruthers[47] suggests, then this would act as a powerful "attractor" for intuitions about responsibility. In particular, it may lead to the intuition that only decisions that are consciously arrived at or that are (or would be) consciously endorsed are truly the subject's own, as we saw in Section 3. And this in turn will generate the intuition that decisions that *aren't* reflectively arrived at and *aren't* consciously endorsed are ones for which the agent can't be held responsible. If this diagnosis is correct, then the ultimate source of the intuitions in question may not lie in the structure of our concept of responsibility, nor in the nature of responsibility itself, but rather in one of the heuristic (and demonstrably false) assumptions built into the mindreading faculty. At the very least, the possibility that this is so should be enough to deprive the intuitions in question of their probative force.

This isn't to say, of course, that there might not be other ways of saving the intuition that Klepto isn't really responsible for his thievery, or that Harry isn't really responsible for injuring his wife.[48] But it does mean that

---

[45] E.g., M. Hauser, *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong* (New York: Ecco, 2006).

[46] B. Scholl, 'Object Persistence in Philosophy and Psychology', *Mind and Language* 22 (2007), pp. 563-91.

[47] Carruthers, 'Cartesian Epistemology: Is the Theory of the Self-Transparent Mind Innate?'; Carruthers, *The Opacity of Mind*.

[48] One might argue, for example, that neither Klepto nor Harry "could have done otherwise", and it is for *this* reason that neither is responsible. However, such a claim is not only

explaining such intuitions should no longer be taken as one of the main goals of theorizing about responsibility. And it may (or may not) turn out that when theories of responsibility have been developed in such a way as to be free of any commitment to the existence of conscious attitudes, the best of them will entail that Klepto and Harry are fully responsible for their actions.

5. Concluding Thoughts: Responsibility without Conscious Attitudes

Our goal in this paper has been to motivate and begin exploring a conditional question: if it should turn out that there are no conscious attitudes, then what implications would this have for theories of moral responsibility? We take ourselves to have made a case for one negative implication over the course of Section 3: Real Self accounts would no longer be viable (or not as currently motivated). We leave it to others, or to another occasion, to consider whether there are other extant theories of responsibility that might suffer a similar fate. We should stress, however, that it won't be enough just to examine whether a given theory is formulated in such a way as to *refer* to conscious attitudes. Rather, one would need to look closely at the underlying motivations of the approach, and to scrutinize the sorts of examples that are used to lend it support. (This is what we have attempted to do in Section 3 with respect to Real Self views.)

Note, moreover, that there will always be two distinct ways for theorists to respond, if they should be forced to recognize that their theories depend upon conscious attitudes, and if it should turn out that there aren't any such things. One would be to give up on the theory of responsibility in question, and to begin seeking another. But the other option would be to insist that the theory in question had been the best available. Thus, theorists would be free to claim that since their theories presuppose the existence conscious attitudes, the proper way to respond in face of the discovery that there are none is to become a nihilist or skeptic with regard to moral responsibility. Such a position wouldn't obviously be absurd.[49]

---

controversial, but full of difficulties of its own. Indeed, part of the appeal of Real Self theories lay in their ability to capture Klepto's and Harry's non-responsibility without appeal to any controversial condition. And we again stress that our aim has not been to show that Klepto and Harry must necessarily be responsible for what they do, but rather to motivate critically assessing assumptions about consciousness to thinking about responsibility.

[49] For discussion of the possibility of living without responsibility, see S. Smilansky, *Free Will and Illusion* (Oxford: Oxford University Press, 2000) and D. Pereboom, *Living Without Free Will* (Cambridge: Cambridge University Press, 2001).

In closing we want briefly to consider some of the positive directions in which theorizing might proceed if one were to accept that there are no conscious attitudes. We will distinguish two broad possibilities. One would be to espouse what could be called "the simple view". This would claim that any actions that issue in the normal sort of way from the attitudes of the agent are ones for which he is responsible. This would still enable us to make important distinctions between involuntary reflex movements and actions done only accidentally or mistakenly, on the one hand, and intentional and deliberate actions, on the other. But it would draw no distinctions among the many ways that attitudes can normally be caused or do their causing. Hence all three of Harry, the second Oedipus, and Klepto would turn out to be fully responsible for their actions, on this kind of account.

The other possibility would be to seek some *other* basis for distinguishing between mental states that belong to one's real self and those that don't. On this approach one would continue to look for some restricted locus of responsibility within the minds of subjects, in such a way that one might deny that some or all of Harry, Oedipus, and Klepto are responsible. But the criterion used, together with any examples held to motivate it, would have to be carefully examined to demonstrate their independence of any commitment to the existence of conscious attitudes. This would, we think, be a challenging task. But we don't claim that it can't be done.[50]

[50] We are grateful to an anonymous referee for this journal for insightful comments on an earlier draft.