Benjamin Kozuch
& Shaun Nichols

Awareness of Unawareness

Folk Psychology and Introspective Transparency

Abstract: *A tradition of work in cognitive science indicates that much of our mental lives is not available to introspection (e.g. Nisbett and Wilson, 1977; Gopnik, 1993; Wegner, 2002). Though the researchers often present these results as surprising, little has been done to explore the degree to which people presume introspective access to their mental events. In this paper, we distinguish two dimensions of introspective access: (i) the power of access, i.e. whether people believe they can unfailingly or only typically introspect mental events; and (ii) the domain of access, i.e. what types of mental events people believe they are able to introspect. We report four experiments carried out to discover where lay beliefs about introspection fall on these dimensions. In our experiments, people did not presume universal introspective access, but they did overestimate the amount of access they actually have, particularly in the case of decisions.*

1. Introduction

It is commonplace in cognitive science that much of our mental lives is not accessible to introspection. It is almost equally commonplace that this result is an affront to common sense, that people presume their current mental events are largely transparent to introspection. Indeed, much of the interest of the attack on introspection has derived from the presupposition that the limits on introspection are *surprising*. In a recent article in the *Journal of Consciousness Studies*, Peter Carruthers has offered the most explicit statement on the matter (2008).

Correspondence:
Shaun Nichols, Department of Philosophy, University of Arizona
Email: sbn@email.arizona.edu

Journal of Consciousness Studies, **18**, No. 11–12, 2011, pp. ??–??

He maintains that a belief in the transparency of the mind is both species-universal and innate. However the extent to which people believe that their minds are transparent has not been systematically tested.

The current paper is a preliminary effort at empirically investigating the folk psychology of transparency. The immediate provocation for our efforts is Carruthers' recent paper on the topic. But Carruthers is making explicit what has been widely presumed in cognitive science. That provides ample motivation for our empirical endeavours. In addition, understanding the folk psychology of introspection promises to illuminate how we think about ourselves, our actions, and the mind more generally. For example, if people presume introspective access, this might help explain why people believe that their actions are not determined (see, e.g. Nichols, 2004, p. 492).

The plan for the paper is as follows: we will first review some classic results on introspection indicating that we lack access to some of our mental events. We will also look at Carruthers' recent proposal about how to empirically explore the issue of transparency. In section 3, we draw some distinctions in order to formulate specific hypotheses about introspective transparency. In the subsequent section, we present the experiments we have conducted. Our results suggest that, although people do not assume global transparency, they do seem to overestimate the degree of access they actually have. In section 5, we speculate on the reason for this overestimation.

2. The Limits of Introspection

The first thoroughgoing attempt to catalogue the limits of introspection appeared in the Nisbett and Wilson paper 'Telling More Than We Can Know' (1977). In this work, the authors brought together a wealth of studies showing that people make mistakes about their own mental processes. In one representative experiment (Nisbett and Schacter, 1966), subjects were requested to take electric shocks of increasingly high voltage. Prior to the shocks, all subjects were given a pill, which was a placebo. Subjects in the 'Pill Attribution' condition were told that the pill would generate heart palpitations, irregular breathing, and butterflies in the stomach. These are in fact typical symptoms of electric shock. Other subjects (in the 'Shock Attribution' condition) were told that the pill would have effects like numbness in the feet, itchiness, and a slight headache. These are not typical symptoms of electric shock. The researchers predicted that the subjects in the Pill Attribution condition would tolerate higher voltage because they would attribute their symptoms to the pill, and not the shock. Their prediction was met

— this group took an average of four times as much shock. The important feature of this study for present purposes, though, is that these subjects failed to recognize the role their beliefs about the pill had on their behaviour. Nisbett and Wilson write:

> Following his participation in the experiment, each subject in the pill attribution group was interviewed following a Spielberger-type (1962) graded debriefing procedure. (a) Question: 'I notice that you took more shock than average. Why do you suppose you did?' Typical answer: 'Gee, I don't really know… Well, I used to build radios and stuff when I was 13 or 14, and maybe I got used to electric shock.,' (b) Question: 'While you were taking the shock, did you think about the pill at all?' Typical answer: 'No, I was too worried about the shock.' (c) Question: 'Did it occur to you at all that the pill was causing some physical effects?' Typical answer: 'No, like I said, I was too busy worrying about the shock.' In all, only 3 of 12 subjects reported having made the postulated attribution of arousal to the pill. (Nisbett and Wilson, 1977, p. 237)

So, many of these subjects took additional shock because they believed that the symptoms they were experiencing were caused by the pill and not the electricity. But the subjects failed to appreciate that they were even thinking about the pill. Apparently these subjects *lacked access* to those mental events leading up to the decision to take more shock. This evidence indicates that our introspective access is limited — we do not have *universal* access to our mental events.

In a different experiment, Nisbett and Wilson (1977) had subjects memorize lists of word pairs. The researchers expected this to produce associations that would affect subsequent responses, and it did. For example, subjects who had memorized the word pair 'ocean-moon' were more likely to name 'Tide' when asked for the name of a laundry detergent. What was important for Nisbett and Wilson, though, was that the subjects were generally unaware that the memorized word pairs had this effect. Nisbett and Wilson write: 'Despite the fact that nearly all subjects could recall nearly all of the words pairs, subjects almost never mentioned a word pair cue as a reason for giving a particular target response' (*ibid.*, p. 243). This again is evidence that our introspective access is limited, as subjects were apparently unaware of the associative effects of the words they memorized.

Nisbett and Wilson took these limits on introspection to be counter-intuitive, maintaining that people believe themselves to have 'direct access to their own cognitive processes' (*ibid.*, p. 255). Because of this, Nisbett and Wilson felt compelled to offer an account of why we would be 'unaware of our unawareness' (*ibid.*). This is a familiar theme among theorists who argue for limits to introspective access

(e.g. Gopnik, 1993; Wegner, 2002; 2004; Gazzaniga, 1995; 2000). Throughout this literature, there is a presumption that the average person *does* think he has access to his mental events, and as a result an attack on introspective access is usually coupled with an explanation for why we mistakenly think we have such wide introspective access.

Carruthers (2008) provides the clearest, most explicit, articulation of the idea that people presume the mind to be transparent to itself. He maintains that a belief in the 'self-transparency of mind' is probably universal, arguing that it is an evolutionary adaptation. Carruthers presents the self-transparency thesis as a conjunction of two claims:

1. 'If I believe that I am undergoing a given mental event, then so I am' (*ibid.*, p. 30).
2. 'If I am undergoing a given mental event, then I can immediately know that I am' (*ibid.*).

Our interests in this paper are entirely restricted to the second claim. We will take the terminological liberty of retaining the label 'transparency', but restricting its scope to the idea that mental events are introspectively accessible.[1] We will leave aside issues about whether beliefs about current mental states are always true.

Carruthers offers a variety of considerations in favour of the view that introspective transparency is a universal aspect of folk psychology. For example, he claims that this view explains why the doctrine of the transparent mind has dominated western philosophical theorizing about the mind.[2] Carruthers also maintains it would make sense for the folk to believe in transparency because not doing so would vastly complicate the operation of a mind-reading system, causing a

[1] Carruthers calls claim 2 'self-intimation'. We prefer not to use that label because it is often associated with the stronger claim that if I am undergoing a given mental event, then I *do* know that I am undergoing that mental event (Sartre, 1956; Brentano, 1874/1973; for a discussion of such theories, see Vollmer, 1999, chapter 5).

[2] This interpretation of the history of philosophy is not universally accepted. One prominent interpretation of Plato's view of akrasia leaves open the possibility that the akratic individual doesn't know his own reasons for action (Bobonich, 2007). In the Modern era, Leibniz quite explicitly rejects introspective transparency. Famously, he posited 'petite' perceptions, perceptions that do not reach the level of consciousness. More surprisingly, he even seems to think that there are unconscious desires that can influence our behaviour. He suggests that we are always feeling at least a small amount of suffering, even at moments when we believe ourselves to be at ease. These states of unease could influence our behaviour even if they escape consciousness, causing us to choose one option over another even in cases where it seems as we are indifferent to which option we choose (Leibniz, New Essays, pp. 188–9; see also Youpa, 2004). Although Leibniz complicates Carruthers' historical claim, it is plausible that Leibniz's view here is driven by his theoretical commitments elsewhere (see, e.g. Adams, 1994). That is, he might have been promoting a theory that was counter-intuitive even to him.

great loss in its efficiency with no gains in its accuracy (Carruthers, 2008, p. 39).

Although Carruthers canvasses a number of considerations in favour of the claim that all people presume introspective transparency, he maintains that the most direct way to assess the universality claim would come from empirical work in developmental psychology and anthropology (*ibid.*, p. 48). Carruthers even suggests a particular vignette as a representative test for whether people believe the transparency thesis:

> Suppose that Mary is sitting in the next room. She is just now deciding to go to the well for water, but she doesn't know that she is deciding to go to the well for water. Is that possible? (*Ibid.*)

Carruthers (*ibid.*, p. 48, fn. 7) reports unpublished pilot work by Clark Barrett indicating that the Shuar of Ecuadorian Amazonia found this scenario impossible. By contrast, they found the following similar scenario possible:

> Suppose that Mary is sitting in the next room. She is just now deciding to go to the well for water, but John doesn't know that she is deciding to go to the well for water. Is that possible? (*Ibid.*)

These are encouraging results for the idea that people presuppose transparency. Our paper aims to extend these investigations.[3]

3. Some Distinctions

Though it's certainly plausible that people believe that the mind is transparent to itself in *some* sense, it is important to generate more specific proposals. We begin by introducing two different dimensions of the folk psychology of introspection.

First, there are questions about the *power* of introspective access. One important view in philosophy is that simply by virtue of undergoing a mental event one is thereby aware of undergoing that mental event (Sartre, 1956; Brentano, 1874/1973; see also Vollmer, 1999). We might call this automatic access. A weaker view is that, while one is not automatically aware of all of one's current mental events, one always *can* be aware of one's current mental events. We can call this

[3] On certain views, it might appear trivially true that all mental states are at least *potentially* accessible. Searle, for example, holds that 'the notion of an unconscious mental state implies accessibility to consciousness' (1992, p. 152). For the purposes of this paper, we are setting aside this view, which is not widely accepted in cognitive science. But it's worth noting that, even on Searle's view, there remain questions as to how reliable the folk think this access is, and whether they think it is more reliable in certain cases rather than others. These are the kinds of questions we are trying answer in this paper.

unrestricted access. An even weaker view is that, while access to one's mental events isn't unrestricted, it is characteristic. On such a view, if I am undergoing a given mental event, then typically I can immediately know that I am. It remains unclear which of these (if any) captures how the folk think about the power of introspective access.

The second dimension to explore concerns which aspects of the mind are thought to be introspectively available. This is a question about the *domain* of introspectively available mental events. Here we rely on a familiar kind of distinction between *states* and *processes*. Mental events such as feeling happy, thinking that it's noon, and intending to go bowling will all count as mental states. Mental processes typically involve relations between mental states. For example, mental processes would include: being made angry by thinking about irresponsible bankers, deciding to get a sandwich because one is hungry, and forming the belief that someone is untrustworthy as a result of discovering him lying.[4]

Appealing to this distinction, there are a number of different views people might hold about introspection: people might presume that they have introspective access to both states and processes; they might presume access to mental states but not processes; or they might presume access only to particular classes of mental states or processes. For instance, it might turn out that people think that they have access to the states implicated in decision making, but not to the states implicated in memory formation.

Before moving on to our experiments on the folk psychology of transparency, we need to return to the psychological work on the limits of introspection. Nisbett and Wilson seem to allow that people really do have access to at least some mental *states*. For instance, they allow that 'an individual may know that he was or was not attending to a particular stimulus or that he was or was not pursuing a particular intention' (Nisbett and Wilson, 1977, p. 256). It is mental *processes* to which people are said to completely lack access (*ibid.*, pp. 255–6). In their textbook, Nisbett and Ross elaborate on this proposal. They describe Aristotelian and Newtonian accounts of gravity and go on to write, 'None of the accounts is an observation of a causal process, since causal processes cannot be observed; instead they are theory-

[4] This way of dividing up mental events is similar to the 'content-process' distinction (Nisbett and Wilson, 1977, pp. 255–6; Nisbett and Ross, 1980; Rakover, 1983). According to this distinction, contents include mental events such as one's sensations, memories, emotions, and plans, whereas processes consist of causal relations between contents. Flanagan draws a related distinction between propositional attitudes and the causes of the attitudes (Flanagan, 2004, pp. 193–5).

guided inferences… A mental process, that is, the means by which one mental event influences another, cannot be observed but only inferred' (Nisbett and Ross, 1980, p. 205; *cf.* Hume, 1739/1963). Thus, Nisbett and Ross seem to be saying that people can't introspect mental processes because that would involve perceiving causal processes, which by their nature are unobservable. Given the context — psychological evidence on introspective limits — this is a rather peculiar comment; we don't need experiments to show that people can't observe the unobservable![5] So we need to say more about the proposal that we lack access to mental processes.

One possibility is that, while mental processes are in principle unobservable, we have a perceptual or cognitive illusion which makes it seem to us that we observe our mental states causally interacting. There is some reason to suspect that we have an experience as of causal process when we perceive colliding billiard balls (see, e.g. Schlottmann and Shanks, 1992). Plausibly, the deflection of a billiard ball is something that appears *causally necessitated*. However, when we turn to the central case of making a decision, it doesn't feel like a case of causal necessitation. Terry Horgan makes this point in discussing the phenomenology of action:

> Although often one does experience certain conscious reasons (e.g., occurrent beliefs, occurrent wishes, etc.) as playing a state-causal role in relation to one's action, this role is experienced as one's being inclined by those reasons to perform the given action; the role is not experienced as one's action being necessitated by those reasons. (Horgan, 2007, p. 9; also Holton, 2006; Vollmer, 1999)[6]

All of this suggests that we need to be careful about the suggestion that people presume introspective access to their mental processes. Nisbett and colleagues contend that people wrongly assume such access, but what is the nature of this putative folk assumption? We offer a catalogue of possibilities for how the folk might think about access to their mental processes involved in decision making:

i. *Direct perception*. We directly perceive causal processes among our mental states. I perceive the mental states *causing*

---

[5] See Nahmias (2002; 2005) for a discussion of Wegner (2002), who sometimes seems to argue in a similar vein. It's possible that Nisbett and Wilson have something more substantive in mind than is suggested by this passage of Nisbett and Ross. In any case, our aim here is not to challenge Nisbett and Wilson, or Wegner, but rather to get clear on the various options.

[6] From the context this quote appears in, it is clear that 'action' should be understood broadly, to involve mental actions such as making decisions or forming intentions.

my intention, much as I perceive the cue ball causing the 8 ball to move.

ii. *Agent causation*. The *agent* causes the decision, but she does it because of some reason, and she can tell which reason it is. I can detect which mental states are causally implicated in my decisions, because I know which reasons I *chose* to act on.

iii. *Inference over introspected states*. We have introspective access to the mental states that are involved in decision making, and we make *inferences* about which of those accessible mental states caused our decisions. I can detect which mental states are occurrent when I make a decision, and I make inferences about which of those states caused my decision.

iv. *Inference alone*. We have no introspective access to states or processes; all of our judgments about mental processes are entirely based on inferences. I infer the causes of my decision based on other inferences about which states I have.[7]

As we see it, it is phenomenologically implausible that the folk embrace the direct perception (view [i]) of mental causation (at least for most mental processes)[8] (Horgan, 2007). At the same time, it's implausible that people regard their introspective abilities as so weak that one must rely entirely on inference (view [iv]) to know anything (at all) about one's own mind. Neither of the other options can be disregarded, however: it might be that people think that when they make choices they have access to the reasons for which they chose; or it might be that people think that they can make good inferences about how their (introspectively accessible) mental states are related. It could also be that people think about access to decisions in different ways in different circumstances. Our experiments below will not distinguish between these different pictures ([ii] and [iii]) of the nature of the access. But at least we can see a space for views about access to processes that are not implausible to attribute to the folk. Now we will turn to explore how people think about the power and domain sensitivity of introspection.

---

[7] Although previous theorists have not distinguished these views when discussing the folk psychology of transparency, some of these views are embraced as apt theories of introspection itself. For instance, (ii) is likely embraced by O'Connor (1995), (iii) is likely Nisbett and Wilson's (1977) view, and (iv) is Carruthers' view (2010).

[8] The direct perception view looks slightly more plausible for select mental processes. Belief formation, for example, may sometimes appear to be causally necessitated. A perception of Bob might bring about, spontaneously and involuntarily, a belief that 'Bob is here'. However, such a view seems implausible in the case of mental events such as decisions, which is what we are mainly concerned with in this paper.

4. Testing the Assumption of Transparency

As indicated in the previous section, we think it plausible that the folk believe the mind to be transparent to itself — at least to a degree. However, it remains to be established whether they take themselves to have access just to mental states, or to mental processes as well as mental states. In addition, there is a question as to whether they think this access to mental states and/or processes is unfailing, merely typical, or something less. Finally, it could be that their belief about access varies according to the kind of state or process under consideration. Given that so much research focuses on limited introspection of *decision making*, we will focus several studies on whether people are more inclined to expect transparency in the case of decisions, relative to other mental events.

All of the experiments we will report are survey-style experiments, in which we collect people's explicit responses. There is some worry about the use of such explicit measures.[9] Explicit measures have a number of familiar shortcomings. For example, with explicit measures, participants sometimes give distorted responses driven by their expectations of what the experimenter is looking for.[10] By contrast, with *implicit* measures, like reaction time and looking time measures, there is much less risk of experimenter demand. We acknowledge that implicit measures are important and not displaced by explicit measures. But that hardly renders explicit measures useless. Indeed, research using explicit measures has often guided subsequent work using implicit measures.[11] Furthermore, in many cases, the implicit measures fully corroborate findings arrived at by explicit measures (compare, e.g. Tremoulet and Feldman, 2000, and Gao *et al.*, 2009). More importantly, implicit measures are notoriously difficult to interpret,[12] and, as a result, evidence from implicit measures is typically most convincing when combined with complementary evidence from explicit measures. Finally, in the case at hand, part of the issue is

[9] Peter Carruthers and Brian Scholl have both raised this concern in personal communication.

[10] With explicit measures, it's difficult to be sure that such experimenter demand isn't playing a role. However, in all of the studies we conducted, we are investigating *comparative* responses, and in each experiment, we find significant differences between conditions. The critical question in each case is: 'Why is there a difference?' And this question demands an answer even if experimenter demand is influencing responses.

[11] For example, the classic false belief task (Wimmer and Perner, 1983) is an explicit measure, and it has been the basis for numerous implicit measures (e.g. Clements and Perner, 1994; Onishi and Baillargeon, 2005).

[12] Just to take one example, classic looking time tasks (e.g. Wynn, 1992) seemed to show that infants engaged in addition, but many theorists reject this rich interpretation in favour of less intellectually impressive explanations (e.g. Mix *et al.*, 2002).

whether people think that it is always *possible* to access your own mental states. It is difficult to investigate this without using explicit report. So while we hope to see implicit measures used in this domain, we think it entirely appropriate to start exploring this issue via explicit measures, as we do below.

Before moving on to our experiments, we should note one more thing. Even if the folk have explicit beliefs about what kinds of mental events they do and do not have access to, they almost certainly do not think about it in terms of 'introspective access', or the distinction between mental 'states' and 'processes'. So we couldn't ask subjects whether people 'have introspective access to mental states'. Instead, we couched questions to our participants in more familiar terms.

4.1. Experiment 1: States and Processes

One central question from the foregoing concerns the *domain* of introspective transparency that is presumed by the folk. The most basic question is whether the folk presume access just to mental states, or to both states and processes. Our first experiment aims to investigate this issue.

As we saw in section 2, Carruthers provided a format for studying people's intuitions about whether a person could make a decision but not have access to that decision. In this study, we took Carruthers' format as a starting point for our investigation. Following Carruthers, we framed our first experiment in modal terms. In Carruthers' pilot study, participants were asked whether it is *possible* for Mary not to know that she is deciding to go to the well. Participants in our study were presented with four statements, each of which described a scenario in which a person lacked access to a mental event.[13] For each statement we asked to what degree the participants thought it was possible. We used four basic scenarios, varying only whether it was about a mental state or a mental process. An example of a scenario about lack of access to a mental *state* went as follows: 'John is just now deciding to go outside, and even though he's paying close attention to his thoughts and feelings, he doesn't know that he is deciding to go outside.' The corresponding *process* question (which was presented to the other half of the participants) was exactly the same, except the word 'that' was replaced with 'why'; e.g. 'John is just now deciding to go outside, and even though he's paying close attention to his thoughts and feelings, he doesn't know why he is deciding to go outside.' Another scenario asked about *thoughts*: 'Frank is just now thinking about the

[13] $N = 32$ undergraduates at the University of Arizona.

beach, and even though he's paying close attention to his thoughts and feelings, he doesn't know why[/that] he is thinking about the beach.' The two other scenarios asked about feeling happy and feeling an urge. For each scenario, the participants were asked to rate their agreement with the statement 'it's possible that this really could happen' on a 1 (strongly disagree)–7 (strongly agree) scale.

There was no clear evidence that participants thought it impossible to lack access to mental states. On the mental state questions, the mean response was close to the middle of the scale ($M = 4.09$). The situation was much clearer for mental *processes*. Here participants tended to agree that it really could happen that the person might not know why he is undergoing the process ($M = 5.34$), and this differed significantly from responses to whether it's possible to lack access to mental states (t (31) = -3.147, $p<0.01$, two tailed).[14]

Thus, it is far from obvious that people think it's impossible to lack introspective access to current mental events. But we do not want to draw any strong inferences from the middling responses participants gave to the mental states question. What is important is the *difference* between these responses and those for the closely matched cases involving mental processes. That comparison suggests that people are more likely to think it is possible to be ignorant of the mental processes leading to one's decisions, thoughts, feelings, and urges.

As noted earlier, empirical work on introspective limits has largely focused on our limited access to the process of decision making. The results from Experiment 1 indicate that people allow that it's possible to be unaware of the reasons for one's decision. However, it remains to be seen whether decision making is regarded as *typically* available to introspection. That is, in our experiment, people thought it possible to lack access to their mental processes. But that says nothing about whether they think this is typical or aberrant. We take this up in our next experiment.

[14] To break this down by question type, for *thinking*-state the mean response (standard deviation in parentheses) for state was 3.88 (2.0), for thinking-process $M = 5.53$ (0.99). For decision-state, $M = 3.94$ (1.85); for decision-process, $M = 5.2$ (1.37). For feeling-state $M = 4.07$ (1.87); for feeling-process, $M = 5.24$ (1.75). For urge-state, $M = 4.53$ (2.07); for urge-process, $M = 5.41$ (1.84).

4.2. Experiment 2: Decisions and Urges

For our second experiment, we wanted to explore whether people thought it more likely that one would have access to the processes eventuating in a decision as compared to the processes eventuating in a different kind of mental event — *urges*.[15] Participants were presented with the following two sentences (counterbalanced for order), and asked to rate how strongly they agreed with them (again on a 1–7 scale).[16]

> 'When I am *making a decision* about what to do (for example, deciding whether to go swimming), if I pay attention to my thought processes, I can usually see what leads me to make the decision I do.'
>
> 'When I am *feeling an urge* to do something (for example, feeling an urge to go swimming), if I pay attention to my thought processes, I can usually see what leads me to feel the urge I do.'

Overall, participants showed some agreement with both claims, but the mean level of agreement for the decision sentence ($M = 5.47$, S.D. 1.36) was higher than that for the urge sentence ($M = 4.69$, S.D. 1.35). This difference was statistically significant (t (98) = 2.883, $p<0.01$). That is, participants showed significantly greater agreement for the claim that they are usually able to know the causes of their decisions as opposed to urges.

4.3. Experiment 3: Decision and Association

For our next experiment, we wanted to directly explore people's expectations about classic experiments on the limits of introspection. Nisbett and Wilson recount numerous studies in which people fail to appreciate various cognitive influences on their behaviour. We wanted

---

[15] To ensure that subjects understand the difference between urge and decision, we ran a study in which we asked participants ($N = 43$ undergraduates at the University of Arizona) to briefly explain the basic difference between an urge and a decision. The answers were coded for adequacy independently by two raters. Inter-rater agreement was high (93%), and disagreements were resolved by discussion. Participants showed quite good comprehension of the distinction. Over 90% of the explanations were adequate. Here are a few representative explanations: 'An urge is an inkling to do something. It is typically based on feelings in the here and now. It is not related to weighing the possible outcomes. A decision accounts for the possible outcomes and weighs the results'; 'An urge is a sudden feeling of want, that suddenly something becomes necessary to do. A decision is a "coming to terms" between two or more choices'; 'An urge can be something spontaneous that you'll have every once in a while. Decision is something that is thought about prior to doing.'

[16] $N = 99$ undergraduates at the University of Arizona.

to see whether people *expect* that we should be aware of those influences. We focused on the two classic experiments that were described earlier: the placebo/shock experiment, and the 'Tide' association experiment. Our study was done between participants: each participant was told about one of these experiments.[17] In both of these classic experiments, the original subjects showed unawareness of a critical influence on their mental processes. In the present experiment, however, we did not tell our participants of the subjects' lack of awareness; rather, this is what we asked about. That is, we asked the participants if they thought the original subjects would have been aware of the critical influence on their behaviour.

Participants were told that we wanted to know their thoughts about an experiment that was done years ago. They were then presented with a description of one of the two experiments, but the description contained nothing about whether the subjects were aware of the influencing factors. Those in the placebo-study condition were given the following description of the experiment:

> Researchers asked subjects to take a series of shocks that increased in intensity. Before they were given the shocks, some subjects were administered a pill and told that the pill would lead to heart palpitations, irregular breathing, and butterflies in the stomach. In fact, the pill was phony, but these symptoms are also the most common symptoms experienced by people when undergoing electric shocks. The researchers predicted that the subjects who were told the pill would produce these symptoms (heart palpitations, irregular breathing, etc.) would take more intense shocks than other subjects. The researchers were right. Subjects who were told that the pill would produce heart palpitations, irregular breathing, and butterflies in the stomach accepted far more intense shocks.

Those in the association-study condition were given the following description of that experiment:

> Researchers had subjects memorize a list of word pairs, like 'flower-garden' or 'home-house'. Each subject received 8 such pairs. The researchers predicted some of these word pairs would lead people to make associations with other words. In particular, they predicted that subjects who had memorized 'ocean-moon' would be more likely to say

[17] $N = 90$ undergraduates at the University of Arizona.

> 'Tide' when asked to name a laundry detergent. The researchers were right. Subjects who had memorized 'ocean-moon' (along with 7 other word pairs) were more likely to say 'Tide' than other subjects.

Our experiment depends on participants' understanding the experiments, so the first part of the task was a comprehension check. Participants were asked to explain why the researchers made the prediction they did.[18] Since comprehension is required for an informative response to our query, we analysed responses only for those participants who passed the comprehension check.

After the comprehension question, participants were asked whether the subjects in the original experiment would have been aware of the influence that the psychologists predicted: Nisbett and Schacter focused on whether the subjects in the pill-shock condition were aware of attributing their physical symptoms to the pill. Accordingly, in the placebo-case condition we asked participants to indicate agreement (on a scale of 1 [strongly disagree]–7 [strongly agree]) on the following statement:

> These subjects would have been aware that they attributed some of their physical symptoms (e.g. butterflies, irregular breathing, heart palpitations) to the pill.

For the association experiment, Nisbett and Wilson focused on the fact that subjects tended not to realize that the word cues had an effect on their recall. We framed our question accordingly. We asked participants to indicate agreement with the following statement:

> These subjects would have been aware that memorizing the word pair 'ocean-moon' led them to think of 'Tide'.

While participants in our study tended to expect the subjects to be aware of the attribution of the symptoms in the placebo study ($M = 5.23$, S.D. 0.973), they also tended to expect the subjects to be not aware of the word cue's influence on their associative recall ($M = 3.19$, S.D. 1.52). This was a significant difference ($t(43) = 5.59$, $p<0.001$).

[18] In the comprehension check, we asked participants to explain why the experimenters would have predicted that those who had the phony pill would take more intense shocks. Answers to the comprehension check were coded independently by the authors (with high inter-rater agreement). Half of the participants failed the test, which is distressingly high. Perhaps this is not so surprising given that the students are not provided with much incentive to read the description carefully. More importantly, those who *passed* the comprehension check did so by providing an adequate paraphrase of the experiment, which is good evidence that they really did understand the experiment.

These results provide further evidence against the idea that people assume wide-ranging introspective transparency; for people seem to allow that one would likely not know of associative influences on recall. At the same time, the results provide more evidence for the claim that people presume themselves to typically have access to the processes underlying decisions.

4.4. Experiment 4: Rational vs. Non-rational

The previous studies suggest there is something special about decision making processes — they are taken to be more accessible than processes like urge-formation or association. But what makes decisions special? In our previous study, one way in which the two conditions seem to differ is whether or not a *rational inference* was involved in the mental process under consideration: in the placebo study, it appears as if, when subjects decided to take more electric shock, it was because they had reasoned it was the phony pill (and not the shocks) that was causing their symptoms. In contrast, the associative mental process at work in the moon/tide effect lacks any obvious kind of rational inference. Perhaps, then, a key factor at work in our participants' judgments as to whether or not the mental process was accessible is the perceived presence or absence of rational inference. On the basis of this, we decided to run another experiment, one testing the hypothesis that mental processes involving an apparently rational inference will be more likely to be judged as accessible.

For the experiment, half of the participants were told about an experiment in which a factor affects a person's behaviour in a way that is easily interpreted as rational.[19] The other half were told about a (very similar) experiment that affects a person's behaviour in a way that is not easily interpreted as rational. Our experimental set-up draws on recent findings showing that the mere presence of a drawing of an eye can affect people's behaviour in economic games (e.g. Haley and Fessler, 2005).[20] This is something most naturally regarded as a non-rational affect on behaviour. Thus, for the *non-rational* condition, participants were presented with the following description of the original effect:

---

[19] 98 participants were recruited through MTurk, a website hosted by Amazon.com (https://requester.mturk.com/mturk/welcome). Users of the site can fill out surveys for modest compensation. Recent work indicates that survey data gathered through MTurk is as reliable as that gathered through standard psychology pools composed of undergraduates (see Buhrmester, Kwang and Gosling, forthcoming).

[20] We're very grateful to Trevor Kvaran for suggesting this study for our experiment.

> In a recent experiment, researchers had subjects participate in an economic game. Each player was paired with another player, one of them (determined at random) would get $10 and the other would get $0. The person with $10 would then be allowed to decide whether to transfer any of the money to the other player. The paired players were seated individually at computer stations and did not know who they were playing with.
>
> The critical part of the experiment was that for some subjects, the computer happened to have a drawing of an eye at the top of the computer monitor. The researchers found that these subjects gave more money.

Our participants were then asked to indicate their level of agreement (1 = strongly disagree, 6 = strongly agree) with the following statement: 'The subjects were aware that they gave more money because of the eye.'

For the *rational* condition, instead of a drawing of an eye at the top of the monitor, there was a webcam. The presence of a webcam presumably would make the decision to transfer extra money more rational. Everything about the description was the same except for the penultimate sentence, which read:

> The critical part of the experiment was that for some subjects, the computer happened to have a webcam with a circle drawn around it at the top of the computer monitor.

Participants in the rational condition were asked to indicate agreement with the statement: 'The subjects were aware that they gave more money because of the webcam.'

As predicted by our hypothesis, participants in the *rational* condition tended to say that subjects *would* be aware that the webcam affected their behaviour (M = 3.91, S.D. 1.31), and subjects in the *non-rational* condition tended to say that the subjects would *not* be aware that the drawing of the eye affected their behaviour (M = 2.71, S.D. 1.64). This was a significant difference (t (95.139) = 4.026, p<0.001).

In addition, people's explanations for their answers in the rational condition appealed to awareness of a rational process. Here are some representative examples:

> '…when people know that someone else is watching, they are more likely to do what is perceived as "the right thing".'

> 'The power of guilt and the loss of anonymity prompted the subjects to be more generous.'
> 'The subjects decided to give more money because there was a webcam. They probably felt that giving money was a indication of their morals in the experiment and decided to give more money away.'
> 'They probably thought it was possible they were being watched thereby inducing them to give more.'

By contrast, people's explanations for their answers in the non-rational condition tended to explicitly reject awareness of the non-rational process. Here are some representative examples:

> 'It seems unlikely that people were consciously "aware" of the eye, or if they were they probably didn't know what it was for or why it was there, even if it affected their behaviour.'
> 'I don't think it would occur to anyone that a picture of an eye at the top of the page could affect their decisions. They probably didn't even really notice it.'
> 'It was just an icon. Perhaps subconsciously they thought about being watched because of the icon.'
> 'I don't think the participants paid attention to the eye knowingly.'

Thus, our experiment supports the hypothesis mental processes that appear to involve a rational inference will be more likely to be judged as accessible. Given that decisions are typically taken to involve rational inference, this would explain why people regard decision making processes as especially available to introspection.

4.5. Experiment 5: The Initiation of Behaviour — Decision vs. Reaction

Our previous experiments all looked at access to the *reasons* for decisions. We now turn to a rather different issue — the *initiation* of action by decision. This issue shows up most prominently in discussions about the neuropsychological work of Libet (1985). In Libet's famous experiment, participants are told to flex their wrist at will, but to note the exact spot on a clock when they are aware of deciding to flex. Libet finds that the ramping up of activity in the motor cortex precedes the time when people are first aware of their decision to flex. The results themselves have been the subject of intense controversy (e.g. Jack and Robbins, 2004; Mele, 2007; 2009). In particular, Libet interprets his results as showing that action is actually initiated before the

agent is aware that the action has been initiated, but it's far from clear that this is the case (Mele, 2006). Nonetheless, our interests here concern whether Libet's results — as he interprets them — would be contrary to common sense. If so, this would suggest that people presume they have access to their decisions before they initiate their behaviour.

We conducted pilot studies in which we described Libet's experiment and simply asked whether the results were surprising. A clear majority of our pilot subjects denied that the results were surprising. This might be because of hindsight bias (e.g. Hawkins and Hastie, 1990), and it might also be because people didn't really understand the experiment. We designed the current study to circumvent both of these problems.

Each participant was presented with a description of one of two experiments.[21] One group of participants was presented with a description of a simplified version of Libet's experiment. The key portion of the description was as follows:

> Researchers asked each of several subjects to flex their wrist at the time of their choosing. Before the experiment, the researchers had attached a measuring device (an 'EEG') to the top of the subject's head to measure the electrical activity of the part of the brain that causes bodily movement (like the movement of flexing the wrist). They also measured exactly when the person moved his wrist. What they found was that the activity in the brain area that causes bodily movement occurred ½ second before the wrist movement occurred.

Following this description participants saw an image of a detailed timeline accompanied by an additional description of the study.

The other group of participants was presented with a description of an experiment similar in certain respects to Libet's, but in this experiment the behaviour was withdrawing from a picture of a spider. The key portion of the description was as follows:

> Researchers showed each of several subjects a picture of a big hairy spider. This was known to make people 'withdraw' from the picture. That is, people move back slightly when shown these kinds of pictures. Before the experiment, the researchers had attached a measuring device (an 'EEG') to the top of the subject's head to measure the electrical activity of the part of the brain that causes bodily movement (like

[21] $N = 88$ undergraduates at the University of Arizona.

> the movement of withdrawing). They also measured exactly when the person showed the withdrawing movement. What they found was that the activity in the brain area that causes bodily movement occurred ½ second before the withdrawal occurred.

This description was also followed by a timeline and additional description of the study. After the presentation of the experiment, each participant was asked whether, if she had been a subject in the experiment described, she would have known that she was going to flex (or withdraw) at a point that was the same or earlier than the time of the activity in 'the brain area that causes bodily movement'. All participants were asked to explain their answers.

As with the previous experiment, we were only interested in the responses of participants who understood the experiments that we described. As a result, the explanations were coded independently by each author for quality. We could then analyse responses only from participants who passed a criterion of understanding. Unfortunately, however, some of the participants did not even offer explanations of their answers, so we established two different criteria. On the more permissive criterion, we excluded subjects who clearly didn't understand the scenario (23 out of 88), but included all other participants, including those who either did not provide an explanation, or provided a vague explanation. When subjects in this group were asked whether they would know of their decision to flex (or withdraw) before the activity in the motor cortex, people tended to assent in the Libet-style condition (68%) but not in the withdrawal condition (19%). This yielded a significant difference ($\chi^2(1, N = 65) = 13.931$, $p<0.001$). We also analysed the data using a more stringent screening test, excluding participants who provided either no explanation at all or a vague explanation. Using this more stringent criterion of understanding the difference is even more prominent.[22]

The results of this experiment revealed that people are much more likely to think they have access to the initiation of a behaviour via voluntary decision than the initiation of behaviour via reaction to an aversive image. People expect to know about their decision to move before the motion system gets activated. However, this is not a global expectation about the motion system, for people tended to think that withdrawing from an image would be initiated unconsciously.

---

[22] In the Libet-flex condition, 79% said they would be aware before the activity in the motor cortex, whereas no participants said that in the withdrawal condition. This yields a significant difference ($\chi^2(1, N = 32) = 18.209$, $p<0.0001$).

4.6. Status of the Transparency Assumption

In section 3 we distinguished two dimensions for the transparency assumption: power and domain. Our results show contours of both dimensions. The first experiment revealed that people do not embrace an unrestricted, all-powerful, transparency view, as subjects in our first experiment allowed for the possibility of being ignorant of current mental processes. Nonetheless, our subsequent experiments suggested that people do expect that we *typically* have access to certain mental processes. When we look more closely at particular domains, it turns out that *decision-formation* is treated as especially available to introspection, at least as compared to urges and associations. The last experiment suggests further that people also expect that the decision to act is available to consciousness before the initiation of the behaviour.

One general conclusion from these results is that psychologists were right to maintain that many of their effects are counter-intuitive. What makes them counter-intuitive, though, is not that they run against a general, indiscriminate assumption of transparency. Rather, they run against a lay assumption that *decision making* is typically transparent to introspection. Of course, we don't mean to suggest that the only domain for which people assume transparency is decision making. But we do mean to suggest — or rather insist — that to adequately characterize views about introspective transparency, one needs to specify the domain under consideration.

Before moving to the next section, we should acknowledge an obvious and important limitation of our data — our participants were western undergraduates, many of whom probably have had exposure to the ideas of Freud and in some cases even cognitive science. As a result, it remains open that different results would emerge if our tasks were conducted in different populations. We are certainly interested in knowing whether the results generalize to other populations. Nonetheless, the results here provide a first step towards understanding the folk psychology of introspection. In addition, we take the most important finding here to be *comparative*. While our participants were happy enough to allow lack of access in many circumstances, they tended to expect transparency in cases of *decision making*. This difference demands explanation.

5. Acquisition

Our evidence does not support the view that people embrace an unrestricted form of the transparency assumption. In our studies, people

think it's *possible* for a mental event to occur without knowing why it is occurring. Furthermore, for several kinds of mental events — urges, associations, and withdrawal initiations — it seems like people aren't much inclined to believe that one typically has access to the processes that result in these. However, this was not the case for decisions. In our experiments, people did tend to think that they typically have access to the reasons for a current decision. This assumption, though much weaker than might have been expected, still exaggerates our introspective abilities, as a great deal of evidence shows that we lack access to influences on our current decision making (e.g. Nisbett and Wilson, 1977; Bargh, 1997). In our studies, the placebo-shock experiment provides the most direct indication of an exaggerated sense of introspective acuity. In that experiment, people wrongly predicted that the subject would have been aware of the influence of his belief about the pill. The question before us now is *why*? Why do people have an inflated sense of introspective access to decision making?

Carruthers has a bold theory for why people presume transparency (2009). On his view, the assumption of introspective transparency is an innate, adapted feature of folk psychology. We want to offer an explanation that does not invoke an innate belief in transparency. But we do want to acknowledge that people overestimate the power they have to access their decision making. Why do people make this overestimation? Recent work on explanation can provide the beginnings of an account. It is commonplace that people are sometimes overconfident about their understanding of various phenomena (see Yates *et al.*, 1997; 1998). That provides little help, though, until we know something about why and when people have an inflated sense of their understanding of the phenomena. In an important article, Leonid Rozenblit and Frank Keil (2002) argue that there are certain features that makes some domains especially likely to provoke an exaggerated sense of understanding, or what they call an 'illusion of explanatory depth'.

As part of a series of experiments, Rozenblit and Keil presented participants with several devices, e.g. a zipper, a speedometer, a flush toilet. After a training session in rating one's 'level of understanding',[23] participants were asked to rate their level of understanding of

[23] In the training session, subjects were presented with examples of different levels of understanding (on a 7-point scale) one might have for a device such as a crossbow. Someone with level 1 knowledge, for example, 'might really only know what a crossbow looks like and what it does — shoots arrows', whereas someone with level 4 knowledge would also know less superficial details, such as the fact that the crossbow 'gets more power than a normal bow and arrow because it allows you to pull the string back extra hard and trap it

how the device worked. These initial ratings tended to be quite high. After rating their understanding of several devices, participants were asked to give a detailed causal explanation for one of the devices, and then asked to rate their level of understanding again. This was repeated for a total of four devices. What Rozenblit and Keil found was that after trying to provide a detailed causal explanation for the device, people rated their level of knowledge as significantly lower than they had previously. Participants, it seems, had an illusory sense of explanatory knowledge. Nor was this simply a general overconfidence: In a follow-up study, Rozenblit and Keil asked subjects to rate their level of understanding for various procedures (e.g. how to bake cookies from scratch, how to tie a bow tie). As in the earlier study, after rating their knowledge level, participants were asked to describe the procedure in a step-by-step manner, and, as in the earlier study, they were then asked to rate their knowledge level again. Strikingly, there was no drop in their ratings of knowledge level for these cases. Parallel studies were also run for knowledge of narratives (movie plots) and facts (state capitals). Subjects in the study involving narratives had no significant drop in their knowledge ratings following their explanation of the movie plots. In the case of facts, there was somewhat of a drop in their knowledge ratings, but it was much less than in the case of devices.

Rozenblit and Keil found that people exaggerated their understanding of devices in a way they do not with procedures, narratives, or facts. Rozenblit and Keil have a promising proposal for why there is this difference — it is, they suggest, because of the 'transparency' of the devices they used in their studies: 'The prominence of visible, transparent mechanisms may fool people into believing that they have understood, and have successfully represented, what they have merely seen' (Rozenblit and Keil, 2002, pp. 552–4). The devices in their studies — toilets, locks, and zippers — are systems composed of discrete and salient parts, and this plausibly contributes to the illusion of explanatory depth. They write, 'The more one is aware of discrete, easy to imagine parts of a system, the more one may be inclined to attribute deep causal knowledge of a system to oneself' (*ibid.*, p. 538).

Rozenblit and Keil also argue that the nature of explanation itself might lead to overconfidence, because 'self-testing one's knowledge of explanations is difficult' (*ibid.*, p. 523). The idea here is that, when attempting to explain how something works, the person explaining

there' (Rozenblit and Keil, 2002, p. 527). Subjects were then asked to use these examples as a guide when rating their own levels of understanding of those devices with which they were presented as a part of the experiment.

may lack any feedback as to whether they have arrived at a complete explanation or not. Rozenblit and Keil suggest that when giving explanations, 'one usually has little idea what the final explanation will look like' (*ibid.*). Contrast this with a procedure such as logging on the internet. In this case, it is obvious if and when one has succeeded.

Turning now to the 'device' of the mind, we suggest that decision making is poised to generate an illusion of explanatory depth. In introspection, we find discrete and salient states (e.g. thoughts and intentions) that plausibly reflect critical causal factors in decision making.[24] Appealing to such states yields some apparent success in analysing the operation of our own mind. Analogously, the people in Rozenblit and Keil's studies had *some* success in explaining how zippers and toilets work. In addition, when we try to explain how we made a decision, there will be no way for us to gauge if and when we have arrived at a complete explanation. Indeed, it's likely that people have little idea what a complete causal explanation of an episode of decision making looks like.[25] And finally, our explanations will rarely, perhaps never, be disconfirmed (Nisbett and Wilson, 1977, p. 256; see also Levin *et al.*, 2000).[26] Given all this — the recognition of discrete causal states in decision making, the pattern of apparent success in explanation, a lack of clear end-state for the explanation, and an absence of disconfirmation — it is hardly surprising that we exaggerate our understanding of our own decision making. Our exaggerated sense of our introspective access to our decision making, on this proposal, is an illusion of explanatory depth.

6. Conclusion

A large body of work in cognitive science has shown people to lack access to some of their own mental events. While these results appear surprising, no empirical work had been done to discover to what degree the folk actually assume such access. This paper is meant to begin to fill this gap. In our experiments, we found that people do not

---

[24] This way of putting it is somewhat controversial since some reject altogether the introspective access to current states (e.g. Carruthers, 2009). But even if there is no introspective access, the point might be retained by appealing to an *illusion* of discrete and salient states that seem to be critical causal factors in decision making.

[25] This is especially plausible given that we are very far from having a complete causal explanation of decision making even in cognitive science.

[26] In this respect, decision making seems to be different from memory retrieval. The attempt to explain a current decision rarely results in abject failure; by contrast memory lapses are familiar, and they are made phenomenologically salient in tip-of-the-tongue cases.

assume a strong from of introspective transparency, as they think it possible to lack access to some of their mental events. However, they do take instances of decision-formation to be typically available to introspection, at least more so than mental events like urges or associations. One of our main conclusions from these studies is that the cognitive science of introspection is counter-intuitive, not because it shows people to lack access to mental events *tout court*, but because it shows them to lack access to the events underlying decision making.

This investigation of the folk psychology of introspection leaves several questions for future research. We found that people accorded special introspective status to instances of decision-formation; future investigations could look to see whether other domains of mental life are also regarded as typically transparent to introspection. For example, one might wonder whether the folk take mental state *attitudes* (e.g. desiring, hoping, believing) to be just as accessible as mental state *contents* (i.e. *what* someone is desiring, hoping, or believing). In addition, it remains unclear how people think about their access to decision making — is it regarded as a form of inference over introspected mental states? Or do people think that they know which mental states they *choose* to act on? Finally, it will be important to see whether our results generalize to other populations. Our studies, then, leave open far more questions than they answer, but we hope that they will lead to further investigations of the folk psychology of introspection.

Acknowledgments

We thank Adam Arico, Peter Carruthers, Brian Fiala, Michael Gill, Keith Hankins, Chris Kahn, Trevor Kvaran, Rachel Schneebaum, John Thrasher, Hannah Tierney, Jen Zamzow, and two anonymous referees for comments on a previous draft. Thanks to Rachana Kamtekar, Houston Smit, and Nick Smith for help on historical views about introspective transparency.

References

Adams, R. (1994) *Leibniz: Determinist, Theist, Idealist*, New York: Oxford University Press.

Bargh, J. (1997) The automaticity of everyday life, in Wyer, R. (ed.) *The Automaticity of Everyday Life: Advances in Social Cognition*, Mahwah, NJ: Lawrence Erlbaum Associates.

Bobonich, C. (2007) Plato on *akrasia* and knowing your own mind, in Bobonich, C. & Destrée, P. (eds.) *Akrasia in Greek Philosophy*, Leiden: Brill.

Brentano, F. (1874/1973) *Psychology from an Empirical Standpoint*, McAlister, L. (ed.), Terrell, D., Rancurello, A. & McAlister, L. (trans.), London: Routledge.

Buhrmester, M., Kwang, T. & Gosling, S. (forthcoming) Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data?, *Psychological Science*.

Carruthers, P. (2008) Cartesian epistemology: Is the theory of the self-transparent mind innate?, *Journal of Consciousness Studies*, **15** (4), pp. 28–53.

Carruthers, P. (2009) How we know our own minds: The relationship between mindreading and metacognition, *Behavioral and Brain Sciences*, **32**, pp. 121–182.

Carruthers, P. (2010) Introspection: Divided and partly eliminated, *Philosophy and Phenomenological Research*, **80**, pp. 76–111.

Clements, W. & Perner, J. (1994) Implicit understanding of belief, *Cognitive Development*, **9**, pp. 377–395.

Descartes, R. (1970) *Philosophical Writings*, Anscombe, E. & Geach, P. (ed. & trans.), Maidenhead: Open University Press.

Flanagan, O. (2004) *The Science of the Mind*, Cambridge, MA: MIT press.

Gao, T., Newman, G. & Scholl, B. (2009) The psychophysics of chasing: A case study in the perception of animacy, *Cognitive Psychology*, **59**, pp. 154–179.

Gazzaniga, M. (1995) Consciousness and the cerebral hemispheres, in Gazzaniga, M. (ed.) *The Cognitive Neurosciences*, Cambridge, MA: MIT Press.

Gazzaniga, M. (2000) Cerebral specialization and inter-hemispheric communication: Does the corpus callosum enable the human condition?, *Brain*, **123**, pp. 1293–1326.

Gopnik, A. (1993) How we know our minds: The illusion of first-person knowledge of intentionality, *Behavioral and Brain Sciences*, **16**, pp. 1–14.

Haley, K. & Fessler, D. (2005) Nobody's watching? Subtle cues affect generosity in anonymous economic game, *Evolution and Human Behavior*, **26**, pp. 245–256.

Hawkins, S. & Hastie, R. (1990) Hindsight: Biased judgements of past events after the outcomes are known, *Psychological Bulletin*, **107**, pp. 311–327.

Horgan, T. (2007) Mental causation and the agent-exclusion problem, *Erkenntnis*, **67**, pp. 183–200.

Hume, D. (1739/1964) *A Treatise of Human Nature*, Oxford: Clarendon Press.

Jack, A. & Robbins, P. (2004) The illusory triumph of machine over mind: Wegner's eliminativism and the real promise of psychology, *Behavioral and Brain Sciences*, **27** (5), p. 665.

Johansson, P., Hall, L., Sikstrom, S., Tarning, B. & Lind, A. (2006) How something can be said about telling more than we can know: On choice blindness and introspection, *Consciousness and Cognition*, **15**, pp. 673–692.

Kripke, S. (1980) *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Levin, D., *et al.* (2000) Change blindness blindness: The metacognitive error of overestimating change-detection ability, *Visual Cognition*, **7**, pp. 397–412.

Libet, B. (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action, *Behavioral and Brain Sciences*, **8**, pp. 529–566.

Libet, B. (1992) The neural time-factor in perception, volition and free will, *Revue de Metaphysique el de Morale*, **2**, pp. 255–272.

Mele, A. (2006) *Free Will and Luck*, Oxford: Oxford University Press.

Mele, A. (2007) Free will: Action theory meets neuroscience, in Lumer, C. & Nannini, S. (eds.) *Intentionality, Deliberation and Autonomy*, Burlington, VT: Ashgate Publishing Co.

Mele, A. (2009) *Effective Intentions: The Power of Conscious Will*, Oxford: Oxford University Press.

Mix, K., Huttenlocher, J. & Levine, S. (2002) Multiple cues for quantification in infancy: Is number one of them?, *Psychological Bulletin*, **128**, pp. 278–294.
Nahmias, E. (2002) When consciousness matters: A critical review of Daniel Wegner's *The Illusion of Conscious Will*, *Philosophical Psychology*, **15**, pp. 527–541.
Nahmias, E. (2005) Agency, authorship, and illusion, *Consciousness and Cognition*, **14**, pp. 771–785.
Nichols, S. (2004) The folk psychology of free will: Fits and starts, *Mind & Language*, **19**, pp. 473–502.
Nisbett, R. & Schacter, D. (1966) Cognitive manipulation of pain, *Journal of Experimental Social Psychology*, **2**, pp. 227–236.
Nisbett, R. & Wilson, T. (1977) Telling more than we can know: Verbal reports on mental processes, *Psychological Review*, **84**, pp. 231–258.
Nisbett, R. & Ross, L. (1980) *Human Inference*, Englewood Cliffs, NJ: Prentice Hall.
O'Connor, T. (1995) Agent causation, in O'Connor, T. (ed.) *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, New York: Oxford University Press.
Onishi, K. & Baillargeon, R. (2005) Do 15-month-old infants understand false beliefs?, *Science*, **308**, pp. 255–258.
Rakover, S. (1983) Hypothesizing from introspections: A model for the role of mental entities in psychological explanation, *Journal for the Theory of Social Behaviour*, **13**, pp. 211–230.
Sartre, J.-P. (1956) *Being and Nothingness*, Barnes, H. (trans.), New York: Washington Square Press.
Schlottmann, A. & Shanks, D. (1992) Evidence for a distance between judged and perceived causality, *Quarterly Journal of Experimental Psychology*, **44A**, pp. 321–342.
Searle, J. (1992) *The Rediscovery of Mind*, Cambridge, MA: MIT Press.
Tremoulet, P. & Feldman, J. (2000) Perception of animacy from the motion of a single object, *Perception*, **29**, pp. 943–951.
Vollmer, F. (1999) *Agent Causality*, Dordrecht: Kluwer Academic Publishers.
Wegner, D. (2002) *The Illusion of Conscious Will*, Cambridge, MA: MIT Press.
Wegner, D. (2004) Précis of 'The Illusion of Conscious Will', *Behavioral and Brain Sciences*, **27**, pp. 649–659.
Wimmer, H. & Perner, J. (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, *Cognition*, **13**, pp. 103–128.
Wynn, K. (1992) Addition and subtraction by human infants, *Nature*, **358**, pp. 749–750.
Yates, J.F., Lee, J.W. & Bush, J.G. (1997) General knowledge overconfidence: Cross-national variations, response style, and 'reality', *Organizational Behavior and Human Decision Processes*, **70**, pp. 87–94.
Yates, J.F., Lee, J.W., Shinotsuka, H., Patalano, A.L. & Sieck, W.R. (1998) Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence, *Organizational Behavior and Human Decision Processes*, **74**, pp. 89–117.
Youpa, A. (2004) Leibniz's ethics, in Zalta, E. (ed.) *The Stanford Encyclopedia of Philosophy*, [Online], http://plato.stanford.edu/entries/leibniz-ethics/

Paper received June 2010; revised January 2011.