Rev.Phil.Psych.
DOI 10.1007/s13164-015-0276-x

The Empirical Case against Infallibilism

T. Parent[1]

Abstract Philosophers and psychologists generally hold that, in light of the empirical data, a subject lacks infallible access to her own mental states. However, while subjects certainly are fallible in some ways, I show that the data fails to discredit that a subject has infallible access to her own occurrent thoughts and judgments. This is argued, first, by revisiting the empirical studies, and carefully scrutinizing what is shown exactly. Second, I argue that if the data were interpreted to rule out all such infallibility, the relevant psychological studies would be self-effacing. For they adopt a methodology where a subject is simply presumed to know her own second-order thoughts and judgments–as if she were infallible about them. After all, what she expresses as her second-order judgment is trusted as accurate without independent evidence — even though such judgments often misrepresent the subject's first-order states. The upshot is that such studies do not discredit all infallibility hypotheses regarding self-attributions of occurrent states.

1 Introduction

It is generally thought that evidence from psychology has debunked *infallibilist* views of self-knowledge, i.e., Cartesian-inspired views where a subject *cannot be wrong* when sincerely reporting her own mental states (thoughts, judgments, experiences, etc.). Now it is, indeed, beyond question that many self-attributions of mental states are false. I would also agree that, in light of the evidence for widespread self-ignorance,[1] it is *possible to doubt*

[1]See, e.g., Dutton and Aron (1974), Tversky & Kahneman (1974); Tversky (1996); Wason and Evans (1975); Nisbett and Wilson (1977); Nisbett and Ross (1980); Rubinoff and Marsh (1980); Gazzaniga (1985, 1995); Garrett and Brooks (1987); Devine (1989); Merikle (1992); Gopnik (1993); Bargh et al. (1996); Wegner and Wheatley (1999); Haidt (2001); Wilson (2002); Pronin et al. (2002); Briñol and Petty (2003); Epely & Gilovich (2005); Lucas and Ball (2005); Johansson et al. (2005, 2006, 2008); Pronin (2008); Bortolotti and Cox (2009); Hall et al. (2010, 2013); Bortolotti (2010), and Bortollotti et al. (2012). An anonymous referee reminds me that there are prominent infallibilists regarding occurrent *phenomenal* states (see e.g., Chalmers 2003; Horgan and Kriegel 2007). But in my own work, the focus is on self-attributions of occurrent *thoughts and judgments*. Infallibilism here seems pretty atypical.

✉ T. Parent
parentt@vt.edu

[1] Department of Philosophy (0126), 220 Stanger Street, Blacksburg, VA 24061, USA

Published online: 14 July 2015

virtually any self-attribution. Still, if the data shows that self-attributions are often wrong, it does not follow that they are most often wrong. In fact, it remains conceivable that many times, the truth of a self-attribution is *nomologically necessitated* by the attribution itself. For instance, one might think that as a matter of psychological law, a second-order thought 'I think that *p*' is composed from the thought '*p*' —meaning that what the second-order thought it is about is literally one of its *parts*. (Cf. Parent 2007, 2013a, ms.) That would make the second-order thought 'infallible' in one sense (although not in the Cartesian sense) of 'infallible.'

Yet suggesting even this milder infallibilism may provoke an 'Incredulous Stare' usually reserved for modal realists. The force of such incredulity is admittedly very powerful. However, my aim is to demonstrate that it is misplaced. The existing psychological studies do not discredit all forms of infalliblism. Not only do the warranted inductions from the evidence fail to rule out infallibilism, but also, for somewhat subtle reasons, the relevant studies would be *self-incriminating* if they licensed such a strong anti-infallibilist stance.

2 What is Evidence-Based Antagonism?

The basic claim in dispute may be called 'evidence-based fallibilism' about self-attributions. Here is a rough way to put such a view:

> (EBF) The psychological evidence shows that self-attributions of mental states are fallible.

In this, a 'self-attribution' is any second-order belief, expressible by the subject by a sentence whose normal form begins with ['I' + Ψ-VERB]—i.e., the concatenation of the first-person pronoun 'I' with a verb expressing a psychological property or relation. But note well: The linguistic expression of a self-attribution I often call a 'self-ascription.' Thus, self-attributions include beliefs expressed by self-ascriptions like 'I feel cold', 'I am thinking', 'I am thinking that water is wet', 'I wish it were summer', and so on.

Evidence-based fallibilism is not merely fallibilism about self-attributions, for (EBF) claims that the empirical evidence has basically *shown* fallibilism.[2] That is stronger than the claim that falliblism is true (and this distinction will matter late in the game). I also do not call (EBF) a kind of 'skepticism.' (EBF) does not merely consist in doubts about infallibility (though such doubts are implied). It is more *condemning* than that. Besides, unlike many skepticisms, the view is *evidence-based*. It is not an idle skeptical stance, based on remote possibilities that no normal person would ever take seriously (deceiving demons and such). Instead, it is explicitly based on strong evidence from psychology. And plenty of people take this sort of stance seriously including Schwitzgebel (2008, 2011), Carruthers (2011) and Kornblith (2002, ch. 4; 2012, 2013).

[2] Terms like 'show' may be unclear, but I shall not demand much of them. At most, I assume that if *p* is shown, then it is irrational to reject *p*. (There are other terms in this work that I don't take the time to define, despite their nebulous meanings. But my goal shall be to apply those terms only in uncontroversial instances.)

Even so, there still remains more than one way to interpret (EBF). For instance, it could be read as 'the evidence shows that *some* self-attributions are fallible.' Yet that is uncontentious—no sophisticated science is needed to know that some self-attributions are false. (I doubt Descartes needed Freud to teach him that people sometimes fool themselves, or don't know what they want.) Nevertheless, (EBF) seems quite *in*correct on a different reading, where it says that the evidence shows that *all* self-attributions are fallible. For there seem to be truistic cases of infallibility, e.g., when one thinks 'I am thinking a thought.' If one manages such a thought at all,[3] then the thought is true—its occurrence suffices for its truth.

Plausibly, then, we should interpret (EBF) as opposing infallibilism about self-attributions of *specific types* of mental states. This results in a fallibilist view specific to a mental type *T*:

> (EBF_T) The psychological evidence shows that all self-attributions of mental tokens of *the type T* are fallible.

Traditionally, values for *T* may include:

T = occurrent phenomenal state
T = occurrent thought
T = occurrent belief

Terminological aside: A 'thought' is any state that hosts a content—whereas a belief is a specific type of thought, one where the subject adopts the 'believing-attitude' toward a content (or as I like to say, the 'alethic pro-attitude'). N.B., I often refer to an occurrent belief as a *judgment*.

Mental states like those above are ones that Descartes' infallibilism supposedly concerned. Accordingly, the evidence-based antagonist says that for at least one of these *T*s, the psychological evidence shows that every self-attribution of a *T*-type stateis fallible.[4] Now I would not oppose the evidence-based fallibilist for most choices of *T*. The only evidence-based fallibilist of concern is one who targets self-attributions of *occurrent thoughts* or of *judgments*, which I shall refer to jointly as 'OTJ self-attributions.' And even then, you would not find me saying that all OTJ self-attributions are infallible, but only those of a certain subclass. However, let us leave open what that class is. (I want to avoid implying that the only evidentially permissible infallibilisms opt for the same class

[3] As discussed in Parent (2013b), I think it is an open question whether thought exists. Ditto with the existence of the self. Still, the point remains that *if* the self thinks 'I am thinking a thought,' then it is invariably true.

[4] There is a serious question about what 'occurrent' means (although this is rarely noticed in the literature). One might have thought that 'occurrent' states are just those mental states that are datable, or occur at some specific time. Yet some dispositional beliefs are possessed only for a short duration, despite being non-occurrent (Bartlett ms.). Suppose you see me wearing my favorite green shirt. That I am wearing a green shirt is your occurrent belief, but it might simultaneously cause a dispositional belief that, e.g., I am not wearing a blue shirt. Further, if you see me change into a black shirt, then the occurrent and dispositional beliefs terminate at the same time. Since the duration of the two beliefs is the same, why does only one count as 'occurrent'? (One hopes that the answer is *not*: 'Because only conscious states can be occurrent'—after all, the word 'conscious' has at least 17, and possibly as many as 40 different meanings; see Lycan 1996, ch. 1; Vimal 2009.) Still, I allow the antagonist the use of the word 'occurrent', as long as I can do the same. (We have *some* grasp of the term, and we can work out details if the need arises.)

as I.) Still, the class would include more than just the truistic self-attributions like 'I am thinking a thought' and its ilk. (A non-truistic case I would include, under certain provisos, is 'I am judging that water is wet;' see Parent ms.) To sharpen all this, let us henceforth call the non-truistic cases 'OTJ* self-attributions.' Then, the only evidence-based fallibilist we need to consider is one we can call the 'antagonist:'

(A) The psychological evidence shows that every OTJ* self-attribution is fallible.

According to our antagonist, the psychological evidence demonstrates the fallibility of any non-truistic, OTJ self-attribution.

Note that (A) is still relatively modest. It even permits believing in the actual truth of all OTJ* self-attributions. It only badmouths the idea that they are *infallibly* true, where the '-ible' in 'infallible' expresses some type of necessity. To make this plain, let K be the set of worlds on which the relevant type of necessity is defined (epistemic, doxastic, nomological, or what have you.). Then, one may explicate (A) as:

(A_K) The psychological evidence shows that no OTJ* self-attribution is true at every world of the kind K.

We shall leave open which worlds are relevant, to allow the antagonist the most latitude. (But assume that K-worlds include our world at least.) Regardless, I eventually argue that if something along the lines of (A_K) is true, then so is the following 'no presumptions' principle:

(NP) The psychological evidence means it is irrational in an SA-testing context to *presume*, of some OTJ* self-attribution, that it is true.[5]

In this, an 'SA-testing context' is a context where the truth of a subject's *S*elf-*A*ttributions are being tested, e.g., the experimental contexts of those studies cited in n.1. As for 'presume', we will need it to be defined only with respect to OTJ* self-attributions, and only with respect to SA-testing contexts. So it will be easiest to give a partial definition of 'presume' in those terms:

(Df) S 'presumes,' of a subject's OTJ* self-attribution, that it is true in an SA-testing context iff S believes without independent evidence, of the self-attribution, that it is true.

Above, 'independent evidence' for an OTJ* self-attribution is (at minimum) evidence that the subject has the thought/judgment self-attributed, where this evidence is probabilistically independent of the subject's own understanding of what she thinks/judges.

Joining (Df) together with (NP), we then arrive at:

[5] Observe I am speaking of presuming *de re* rather than presuming *de dicto*. The former makes for a logically stronger condition; presuming *de re* that $(\exists x)\Phi$ secures the corresponding *de dicto* presumption, but not vice-versa.

(NP_{Df}) The psychological evidence means it is irrational in an SA-testing context to believe without independent evidence, of some OTJ* self-attribution, that it is true.

My suggestion now is that (NP), interpreted as (NP_{Df}), is entailed by (A) when (A) is understood in terms of (A_K). I shall simply assume the entailment for now, and give an argument later, once I am in a better position to do so.

3 Revisiting the Experiments

Thus, we are presently assuming that the evidence-based antagonist holds it is irrational to presume the truth of some OTJ* self-attribution, at least in an SA-testing context, given the evidence from psychology. Although I have yet to argue this, the basic thought is rather intuitive. In a context where the subject's self-attributions are being *tested*, it would be odd if researchers simply presumed the truth of some OTJ* self-attribution (especially if results from other SA-tests *show* that such attributions are often false).[6] Even so, I will demonstrate that under this norm, the relevant psychological studies are *self-incriminating*. That's because the experimenters adopt *exactly those kinds of presumptions*.

To be clear, I myself do not regard the studies as self-incriminating—the experiments indeed show significant limits on self-knowledge. Yet that is because I suspect the experiments (unwittingly) exploit the infallibility of some self-attributions, despite revealing other self-attributions to be quite fallible. (Though again, this paper attempts no positive argument for infalliblism; infallibilism here is only playing defense against a type of evidence-based fallibilism.)[7] Still, we shall see that our *antagonist* should have a serious objection to the experiments, even though they provide the empirical basis for her antagonism. In brief, that's because (NP) is violated in those experiments.

What's more, we shall appreciate that the violation of (NP) will seem like a pseudo-problem. This will indicate that evidence-based antagonism is not only justified in a self-prohibited manner, but also that it must be wrong in some fashion or other, insofar as it insists on (NP). The moral will be that the antagonist is incorrect—the evidence does *not* discredit infallibilism about some OTJ* self-attributions. Such a view may be rightly rejected in the end, but my claim is that our antagonist currently lacks proper grounds for that.

Matters shall be clearest if the issue is related to some actual psychological studies. For this purpose, I have selected a classic one by Nisbett and Wilson (1977), and a more contemporary study by Haidt (2001). (However, the issue is hardly unique to these studies.)[8]) The claim, then, is that (NP) would mean that there is a serious methodological problem for both studies.

[6] Take heed that although (NP) concerns self-attributions of occurrent first-order states, there is no assumption that the self-attributions are themselves occurrent (higher-order) beliefs. This is as it should be—after all, it still would be odd for experimenters just to presume the truth of a subject's higher-order belief in an SA-testing context, even if the belief is a dispositional belief rather than an occurrent one.

[7] Moreover, I grant that one need not take this infallibilist line to see the experiments as legitimate. It is enough to say that they exploit the high reliability some self-attributions. But again, my present aim is *not* to give a positive case for infallibility; it is only to rebut the empirical case against it.

[8] Indeed, the issue seems present in *all* studies that I am aware of. For it concerns any methodology that takes subjects' verbal reports as accurate expressions of occurrent beliefs (even if those beliefs are deemed mistaken, re: the mental states they are *about*). This shall become clearer in later sections.

There are a variety of experiments carried out by Nisbett & Wilson, but in the interest of brevity, I shall just discuss one. In this experiment, 81 male introductory psychology students were provided a list of word pairs, with the instruction to commit them to memory. Some of the pairs were designed to facilitate certain associations with other words. For example, when subsequently asked to 'name a detergent,' the memorized pair 'ocean-moon' was meant to encourage subjects to name 'Tide.' And indeed, when interviewed, such associations were exhibited. Target responses (e.g., 'Tide') doubled from 10 to 20 % when subjects had memorized the list of word-pairs. However, in subsequent interviews, subjects failed to cite the memorized word-pairs as a reason for their answers. Instead, subjects 'focused on some distinctive feature of the target 'Tide is the best-known detergent'…or…'I like the Tide box" (p. 243).[9]

As for Haidt's (2001) experiment, subjects first read a short passage about a brother and sister agreeing to have sex in secret while vacationing. The story adds that she is already on the pill, and that he uses a condom as well. Further, although the two have a fine time, they decide not to do it again. And the experience makes them feel closer by sharing in the secret. Haidt reports that, nonetheless, most people judge that what the siblings did was wrong. Yet they are hard-pressed to say why ('moral dumbfounding'). Haidt reports that his subjects:

> point out the dangers of inbreeding, only to remember that Julie and Mark used two forms of birth control. They argue that Julie and Mark will be hurt, perhaps emotionally, even though the story makes it clear that no harm befell them. Eventually, many people say something like 'I don't know. I can't explain it. I just know it is wrong.' (p. 814)

Here, subjects seem engaged in *post-hoc* rationalization, rather than introspecting the actual mental precursors to their verbal response. In so doing, they portray themselves as having replied because of antecedent beliefs *x*, *y*, and *z*, although the reply was often just a knee-jerk reaction.[10]

The experiments support a similar hypothesis, namely, that subjects are unreliable in identifying why they respond in the way they do. So they apparently provide evidence for some form of antagonism, re: self-attributions.

Even so, consider that subjects expressed their faulty self-reports along the following lines:

(1) [I responded with 'Tide' because] 'I like the Tide box'
(2) 'They [= the memorized word-pairs] didn't affect my responses'

[9] Smith and Miller (1978) claim, contra Nisbett & Wilson, that the suggested responses on the part of the subjects may well be true—if the subject's task was to identify *some* causal influence on why she answered with 'Tide'. But while this possibility strikes me as real, I shall bracket it in what follows.

[10] *Pace* Haidt, it is unclear whether subjects meant to describe the *actual* reasons that lead to their responses. Possibly, they might have seen themselves as *producing* reasons to defend their intuitive judgment. (A similar issue can be raised for Nisbett & Wilson, though it is less obvious in that case.) But I shall ignore this in the discussion above.

Question: Might the Haidt-subject be *right* that her response owes to a belief in the harms of incest? In fact, the belief she reports is not that *in general* incest can result in emotional damage or birth defects. It is rather that *this particular act* of incest between Julie and Mark could result in harm. Yet since that overtly contradicts what is known from the story, the subject is likely confabulating. (Thanks to an anonymous reviewer for pressing me to clarify this.)

(3) [I responded that the siblings did something wrong because] 'It could result in emotional damage or birth defects.'

Crucially, this data may not be of the right type to serve antagonism about OTJ* self-attributions. Prima facie, (3) does not even express a self-attribution of any sort. After all, the subject's explicit response just concerns the siblings' act; it does not concern a mental state of the subject herself. Even so, we can plausibly interpret (3) otherwise, where it is shorthand for:

(3.1) [I answered that the siblings did something wrong because *I believe*] 'It could result in emotional damage or birth defects.'

The insertion of 'I believe' in (3.1) makes plain that the response is meant to communicate her belief about the story (rather than the facts of the story per se). We could concurrently suggest that (3.1) is a conversational implicature of (3) (cf. Grice 1975). But suffice it to say that one is a shortened version of the other, and that this is understood in context.

Seen in this way, (3.1) is of the right type to bear on antagonism about OTJ* self-attributions. But what of the other responses? (1) seems to self-ascribe a preference or *liking*, rather than an occurrent thought or occurrent belief. Even so, let us suppose for discussion's sake that the subject uses the utterance to mean:

(1.1) [I answered with 'Tide' because *I believe*] 'The Tide box is likable'

Here, the utterance would expresss an OTJ* self-attribution. As for (2), however, this less readily fits the mold. For it primarily concerns the absence of a causal relation between the subject's earlier answer and other mental states. In this respect, it does not obviously involve OTJ* self-attributions. It is true that (2) (like all assertions) is an *expression* of an occurrent thought. But it is not *about* the subject's occurrent thought or judgment—which is essential to OTJ* self-attributions.[11] I thus suggest we leave (2) aside, since we can at least work with (1.1) and (3.1).

Now (1.1) and (3.1) concern the causal basis of earlier answers as well. Yet they seem to remain relevant—for in these two cases, the subject identifies her *belief* as causing her earlier answer.[12] (In contrast, the causal influence in (2) was a somewhat non-descript state of memory.) But: Are (1.1) and (3.1) about *occurrent* beliefs? It seems not. In each case, the belief is identified as one from the past, as the cause of the subject's answer to an earlier question (White 1980, p. 106). Yet if so, then ultimately *none* of the actual responses from the experiments justify 'evidence-based' antagonism! That is so, given that the antagonism is restricted to OTJ* self-attributions. (But admittedly, there

[11] The memorized word-pairs may constitute a group of associative beliefs, in which case, the subject's reference to the word-pairs might ultimately be spelled out as some kind of OTJ*self-attribution. However, (2) at most implies that some beliefs or other of that type exist. It does not self-ascribe any of the specific beliefs involved.

[12] There is a question about whether these beliefs are causes vs. reasons (cf. Anscombe 1963; Davidson 1963). Well, the experiments assume that they are at least causal influences, and whether they are also reasons is a question we can leave open. I thus speak of causes above, but this should be read as neutral on the reasons-question.

are other ways of construing subjects' responses, and we shall consider these in the next section.)

To his credit, Carruthers (2011) recognizes this sort of issue, despite his clear antagonistic proclivities.[13] However, he cites other experiments (e.g., Wegner and Wheatley 1999), where the time-lapse is quite short between the state and the self-report.[14] He thus suggests that failures of working memory are not a factor in these cases—so as concerns reliability, verbal reports can be counted *as if* they co-occurred with the reported mental states (2011, p. 341). Yet the evidence here remains that subjects are quite fallible in what they report.

However, Carruthers assumes that *memory loss* would be the reason for any decreased reliability. But this is not clear. Consider that a first-order state might occur without being recorded in memory at all. A stray thought might just have a momentary existence in the global workspace,[15] disappearing without a trace. In such a case, if one subsequently fails to report the thought, it is not so much that memory failed to retrieve a record of it. It is rather that no record was ever stored in the first place. In contrast, as long as a thought is presently occurrent, then introspection at least has a chance at detecting it. Thus, there may be a significant difference in reliability between co-occurrent self-attributions and those that are even slightly delayed.

It is tempting to say that a moment later, any thought will leave some trace. But this is dubious, in the same way that it is dubious to say that every facet of the visual field leaves some impression on memory. I am reminded of an anecdote, where Watson claims to recall well a certain building. Holmes then cleverly asks how many street-facing windows the building has—whereupon Watson is unable to answer. This exemplifies a well-known phenomenon, where one remembers x, without remembering every detail about x. In Watson's case, memory never bothered to commit the exact number or arrangement of the windows. So it is not so much that he forgot these things. Rather, he never registered them in the first place.

Plausibly, some momentary, stray thoughts are like the windows on Watson's building. While a thought is occurrent, one can report it reliably enough, just like Watson can report the number of windows while gazing at the building. One can *attend* to the relevant feature. But once the feature is no longer present, such an attentional act is not possible, and memory will not have recorded every passing detail. In which case, one may therefore be significantly less reliable about what was 'in view'—even if only

[13] Surprisingly, however, Carruthers is not particularly antagonistic toward knowledge of our own sensations: '[Regarding] the set of sensory or sensory-involving states, which include seeing, hearing, feeling, and so on…the model of self-knowledge that I present regards our awareness of these types of state as being relatively unproblematic' (p. xi). But see Schwitzgebel (2008, 2011) for counterevidence on this.

[14] Also, Carruthers (2011, p. 344) claims that Briñol and Petty's (2003) study concerns occurrent judgments, viz., self-assessments of confidence. Yet Briñol & Petty's experiment provides no evidence that those self-assessments are *in error*. I would grant that some amount of self-ignorance is shown, but only self-ignorance of a non-rational influence on an *earlier* judgment. (To be fair, Carruthers' primary interest with Briñol & Petty is to confirm his Bem-style self-perception theory, and it may well do so. Still, Briñol & Petty's data do not show self-ignorance of an *occurrent* state.)

[15] As it turns out, Carruthers (2014) now denies the existence of a cognitive 'workspace', for reasons related to his Bem-style self-perception theory. I cannot properly address these arguments (although see Wu's 2014 response). But my point above does not really depend on the existence of such a workspace. It is just that memory is *selective* with respect to the mind's own operations, just as it is selective with respect to external states-of-affairs.

a moment has passed. Contra Carruthers, then, it would not be memory loss, but rather the *selectivity* of memory that would cause such unreliability.[16]

By the way, there is a second reason why co-occurrent self-attributions might be especially reliable, even compared to those occurring a millisecond after. On the view defended in Parent (2007, 2013; ms.), sometimes the first-order thought is literally an (ineliminable) *part* of the self-attribution. If so, then tokening the second-order judgment necessitates the occurrence of the first-order thought—just like writing 'I am writing that water is wet' necessitates writing 'water is wet'. And in both cases, since the occurrence of the first-order bit is precisely what is contended, the second-order bit is thus infallible.

It is clear, moreover, that this sort of infallibility is possible only if the self-attribution temporally overlaps with the entire first-order state. Any delay between them means that this 'compositional' kind of infallibility cannot occur. So again, delayed self-attributions may well be significantly less reliable.

4 The Argument against Evidence-Based Antagonism

Thus, evidentially speaking, there seems to be no substitute for self-attributions that co-occur (roughly) with the reported states. But unfortunately for the antagonist, the existing experiments all feature some time-lapse. So the point remains that the existing evidence does not demonstrate the unreliability of OTJ* self-attributions.

Nevertheless, we might assume otherwise for discussion's sake. Specifically, with (1.1) and (3.1), let us suppose that the belief being self-ascribed is *both* occurrent and a standing belief from earlier. (The belief is *re-occurrent*; the subject is reaffirming what she judged earlier.) To make this clearer, let us rephrase the subjects' responses as conjunctions, where the first conjunct patently expresses an OTJ* self-attribution. In each case, we can then focus on these bits bearing directly on the antagonistic view. The rephrasings look like this:

> (1.2) I judge the Tide box to be likeable, and this is the same belief that caused me to respond earlier with 'Tide'.
> (3.2) I judge that their act might result in emotional damage or birth defects, and this is the same belief that caused me to respond with 'The siblings did something wrong'.

Granted, such adjustments to the subjects' responses mean we are deviating quite a bit from the raw data. But again, I am doing this out of charity. Besides, it is not entirely implausible that (1.2) and (3.2) just articulate more precisely what was actually said in loose terms.

Imagine now that the Nisbett-Wilson respondent is somewhat repelled by the Tide box, and her verbal response is an *ad hoc* rationale for the earlier answer. Similarly, suppose the Haidt subject is mainly concerned to rationalize her initial snap judgment.

[16] An anonymous reviewer worries that this indicates a quite general problem in the psychologist's aim to test a subject's self-knowledge. While this may be so, I do not want to delve into the matter at this point. However, see section 4.3 below for some pertinent considerations. Also, see the concluding section.

Indeed, since Haidt's story directly contradicts the beliefs the subject self-ascribes, we can easily imagine that she has the contrary belief, and momentarily forgot that fact.

Seen from this angle, the experiments would seem to justify antagonism about (at least some) self-attributions of occurrent beliefs. Whether or not that is so,[17] the point I wish to make now is that the experiments ought to be methodologically problematic by the lights of (NP). For, even though the context counts as an SA-testing context, the experimenters are nonetheless presuming the truth of some OTJ* self-attributions.

The key idea here is that, if subjects are seen as mistaken about their first-order beliefs, then the experimenters must be presuming subjects are right about their *second-order* beliefs.[18] For if subjects do not really have the second-order beliefs they express, then they are not mistaken in the way the experimenters suggest. Yet the evidence for attributing these second-order beliefs just consists in the subjects' expressions of those beliefs. (At least, the studies mentioned no other evidence on this point.) However, that means there is no *independent* evidence for a subject having these second-order beliefs—no evidence that is independent of the subject's own understanding of what she second-order believes (as manifested in her assertion). So there is a violation of (NP).[19]

In developing this line, it will help to use Rosenthal's (2005) distinction between a subject *reporting* a belief and *expressing* a belief (See Rosenthal 2005 ch. 2, §2.) To illustrate, suppose I assert 'I believe that dinner is at six'. Normally, this *reports* a first-order belief that dinner is at six, i.e., it is a representation of me having that belief. Yet the assertion is also an *expression of* a second-order belief. The assertion articulates a belief to the effect that I have the first-order belief.

Similarly, if a Nisbett-Wilson subject utters 'I believe the box is rather likable', the utterance *reports* a first-order belief. Simultaneously, however, the utterance is an *expression of* the subject's second order belief—the utterance stems from a belief about what she first-order believes. Now in the Nisbet-Wilson study, the question concerns the truth of the second-order belief, the belief allegedly expressed by her assertion. That's because it is in doubt whether the subject really has the first-order belief, the one

[17] Since one's beliefs/desires can be inconsistent, it does *not* follow from current suppositions that the subject's self-ascription is false. This is especially clear with Haidt's subject: It may be she has *both* the self-ascribed belief and the opposite one as well. Perhaps her self-ascription has a misleading implicature that she is being consistent in her judgment—regardless, the presence of the one belief is sufficient to make the ascription true, regardless of whatever else she believes. Even so, I shall not rely on this point in what follows.

[18] There is a connection here with Davidson's (1983) view that, in order for a speaker to be interpretable, she must be presumed to have mostly true beliefs. The case of SA-testing is arguably a special case of this: In order to interpret a speaker as being mistaken about her first-order beliefs, we must presume she correctly grasps her second-order beliefs. However, I mention it only for those who are interested—nothing above requires adopting Davidson's views. Cf. also Wright's (1998, p. 17) view that self-ascriptions must have a presumption of truth, in order to make sense of the subject as a thinker/agent. Shoemaker (1968) makes a similar claim as well.

[19] One reader remarked that this point was already made at the end of Bem (1967). But for one, Bem's target is only the cognitive dissonance theorists, such as Festinger (1957). For another, Bem's criticism is that these theorists impose much of their own psychology onto the subject when looking for cognitive dissonance. The present point, however, is that experimenters presume that a subject expresses her own second-order belief, versus a confabulated one. Moreover, the basis of this presumption is not that experimenters assume that *they themselves* reliably express their own second-order beliefs, a reliability which they then project onto subjects. (My suspicion is instead that the experimenters' presumption is based on a need to interpret subjects in a way that makes their utterances relevant to the experiment. Cf. the connection with Davidson et al. in the previous footnote.)

she reports herself as having. Yet in all this, the subject's possession of the second-order belief is *not* in question. That belief is assumed to be present; the only concern is whether it is true. In taking its presence for granted, however, are we presuming that some OTJ* self-attribution is true?

On this matter, consider the initial conjuncts of (1.2) and (3.2):

(1.25) I judge the Tide box to be likeable.
(3.25) I judge that the sex act might result in emotional damage or birth defects.

The experimenters claim that, when uttered by the relevant subjects, (1.25) and (3.25) are false. But notice this alone does not imply that subjects have made a mistake, for they must also believe them to be true. That is, if 'I' is indexed to an appropriate subject, it must also be true that:

(1.3) I believe that I judge the Tide box to be likeable.
(3.3) I believe that I judge that the sex act might result in emotional damage or birth defects.

Yet what establishes that these second-order beliefs exist? Apparently, the assumptions in this are:

(4) If a subject asserts 'I judge that *p*,' then (other things equal) she is expressing her belief that she judges that *p*.
(5) The Nisbet-Wilson subject asserts (1.25).
(6) The Haidt-subject asserts (3.25).

From these premises, it indeed follows that the subjects possess the relevant second-order beliefs (at least when *ceteris* is *paribus*). Accordingly, since experimenters accept the premises, they conclude that the evidence shows that subjects have made a mistake.

But here is the crux of the matter. On what basis is (4) accepted? The principle is not trivial; it remains entirely possible for a subject to make an utterance, yet for her not to be expressing her belief (even if the case looks entirely normal). The subject might misunderstand the terms she's using; she might be making a performance error; she might be lying or sarcastic. Conceivably, she might even be engaged in *second-order confabulation*. Now if the case *looks* normal, it is right and natural for the experimenters to ignore such possibilities (or so we may allow for argument's sake). But what exactly does this take for granted?

When there is no defeating evidence, it takes for granted that the subject indeed has the second-order belief she understands herself as having, an understanding that her assertion is based on. Her assertion thus functions as evidence by which the subject is attributed the second-order belief. But as a result, the evidence for the second-order belief is *not independent* of the subject's own self-understanding. In particular, it is not independent of her understanding of what she second-order believes, where this understanding is reflected in which (second-order) belief she chooses to express. So apparently,

the experimenters are simply presuming the subject is right in her understanding of which second-order beliefs she has. And that violates (NP).[20]

As earlier noted, however, the violation of (NP) does not seem like a real problem. The psychological research still strikes me as entirely legitimate. But the problem *is* real if (NP) is true—and that suggests that (NP) is not true. Moreover, assuming that (NP) is a corollary of the antagonist's view, it follows that such antagonism is also mistaken somewhere, at least when it comes to OTJ* self-attributions.

Before moving on, one might ask if the violation of (NP) *ought* to be seen as a problem. For present purposes, this need not be settled—the antagonist is in trouble regardless. If problem is indeed real, then the empirical justification of the antagonist's own view is vexed. Granted, her view might not be sufficiently self-incriminating to be self-*undermining*. I.e., the difficulty might not be serious enough to *disqualify* the evidence as evidence. Still, it should put a damper on her insistence against any infallibilist hypothesis. After all, her own evidential basis treats some OTJ* self-attributions *as if* they were infallible, as if their truth was guaranteed.[21]

5 Objections and Replies

5.1 The Objection from the Intentional Stance

The first objection grants that the experimenters believe (4) in the SA-testing context, but denies that (4) is merely presumed. That's because experimenters have independent reasons for believing (4), reasons that concern the standard belief-desire explanation of behavior. The idea is that *ceteris paribus* a subject would not have made *that* assertive speech act, as opposed to some other one, unless she really had the second-order belief

[20] In fact, this argument goes through only if the subjects' self-understanding includes some *third*-order belief on what she second-order believes. This is needed since the subject's self-understanding is what is presumed accurate—and (NP) is thereby violated only if it amounts to presuming, *of some OTJ* attribution*, that it is true (where an OTJ* attribution is a *belief*). Thus, to get a violation of (NP), the subject's (presumed accurate) understanding must suffice for a third-order belief about a second-order belief. Still, the third-order belief would typically be non-occurrent, and it need not *wholly* constitute her self-understanding. And the experimenters themselves need not posit a third-order belief in all this. For when experimenters presume the accuracy of her self-understanding, they would have only a *de re* presumption, of the third-order belief, that is true. So the premise is really less contentious than it may initially seem. (Thanks to an anonymous referee for urging further clarify on this matter.)

Nonetheless, Proust (2013) is one who holds that some metacognitive states are wholly non-representational. Yet she thinks they can still guide thought/action by constituting a 'non-intellectualist' kind of *know-how*. I cannot fully respond to this view here. As Proust is aware, however, there is some difficulty in construing such 'know how' as *meta*cognition, if the relevant states represent no cognitive event or process. Proust glosses this by talking of these states as 'being sensitive' to specific cognitive events; yet it is unclear why such sensitivity is not a representational affair. (See Langland-Hassan 2014 for a more detailed presentation of this objection.)

[21] My self-incriminating objection should not be confused with another self-effacing issue (noted by Louise Antony, in conversation). This other issue arises from asking the evidence-based antagonist: How do you know you are an evidence based-antagonist? Ordinarily, her self-reporting as much would be accepted without independent evidence. But that too would be a violation of (NP). By the lights of evidence-based antagonism, then, one is ordinarily not warranted in self-attributing that view! But whether or not this objection is sound, my present point is just that the objection above is a different objection. It concerns whether the antagonist can consistently endorse the *experiments* providing the evidence for antagonism. It is not a point about self-identifying per se as an evidence-based antagonist.

she expresses. So in that respect, the experimenters have abductive grounds for accepting the subject's expressions of her own second-order beliefs (hereafter, grounds for 'trusting the subject').[22]

The relevant belief-desire explanations would be similar to more mundane ones. I have in mind explanations like:

(7) She brought her umbrella, as opposed to leaving it at home, because:

(i) She believes that rain is presently likely; plus
(ii) she desires to stay dry, and
(iii) she believes that bringing her umbrella is a way to stay dry.

And insofar as the explanation is plausible, there are abductive grounds for attributing her the judgment ascribed by (7.i).

As for the subjects of the experiments, the explanation would be along the lines of:

(8) The subject uttered 'I judge the Tide box to be rather likable', as opposed to some other sentence, because:

(i) She believes that she judges the Tide box to be likable; plus
(ii) she desires to express that (second-order) belief, and
(iii) she believes that her utterance is a way to express that (second-order) belief.

(9) The subject uttered 'I judge that the sex-act may result in emotional damage and birth defects', as opposed to some other sentence, because:

(i) She believes that she judges the siblings' act might have those results; plus
(ii) she desires to express that (second-order) belief, and
(iii) she believes that her utterance is a way to express that (second-order) belief.

Here, the explanations partly consist in (8.i) and (9.i). So adopting these explanations means attributing the second-order judgments that are expressed. Thus, in the style of abductive arguments, the experimenters have grounds for trusting the subject.

There is nothing irrational about this. It illustrates what Dennett (1975) famously calls the 'intentional stance,' and countless instances confirm that the strategy works well. Nonetheless, the issue is not whether there are some grounds for trust; it is whether there is *independent evidence* for trust, within an SA-testing context. The abductive grounds, moreover, do not count. The short argument here is that, if they did, then the experimenters would also have 'independent evidence' to trust the subject regarding her *first*-order beliefs. That's because the same type of abductive grounds could justify experimenters in attributing such beliefs. But in an SA-testing context, the subject's grip on her first-order beliefs should remain in doubt. So the abductive grounds from the intentional stance are not 'independent evidence' in an SA-testing context.

[22] 'Trusting' is not an ideal word-choice, especially since it has started accruing philosophical baggage in the literature. But I am unable to think of a better word without inventing nomenclature.

5.2 The Rejoinder from Verbal Expression

We shall approach this point again, by first considering a rejoinder at this point. Conceivably, the abductive grounds might suffice as evidence in one case, but not the other, because the evidence is stronger in one instance. In particular, one might argue that the subject's second-order utterance means she is more likely to have the second-order belief, as compared to the first-order belief. That's because a subject's utterance of 'I believe that *p*' is normally an *expression* of her second-order state, viz., her belief that she believes that *p*. In contrast, it only *reports* that she has the first-order belief that *p*. This difference suggests, moreover, that the self-ascription will normally be caused by the second-order belief. And of course, if it is caused by that belief, then that belief must exist. No such guarantee exists for the reported first-order belief.

But why exactly should we favor the second-order belief as the cause of the utterance? Why not say the subject's utterance 'I judge that *p*' has her first-order belief that *p* as its causal antecedent? In many cases, such a causal supposition may be tenable. Still, the evidence-based antagonist could contend, with some plausibility, that the second-order belief is the more proximate or powerful cause. This is so, given that the utterance is second-order. (After all, if the first-order belief were the dominant cause, one would instead expect the subject to utter '*p*' rather than 'I judge that *p*.')

Thus, says our antagonist, if there is less causal distance or a tighter nomological connection between the utterance and the second-order belief, then the utterance is a more reliable sign of the second-order belief. So the evidence is indeed stronger when the second-order belief is posited, in contrast to positing the first-order belief. And that is why, in the SA-testing context, the evidence-based antagonist can maintain that the experimenters have adequate evidence for trusting the subject on which second-order beliefs she has, but not on which first-order beliefs she has.

Now it is still dubious whether our antagonist has identified *independent* evidence for trusting the subject, but waive that. For one can argue that belief-expression should not count as adequate evidence (whether it is independent or not) at least in an SA-testing context. After all, in such a context, it is clearly inadequate for trusting her expression of a first-order belief. That is so, even though a first-order belief would likely be the more direct cause of *that* type of utterance. Put differently, if the context includes a linguistic expression of '*p*', where a first-order belief is more plausibly the dominant cause, it should still be in doubt whether the subject believes *p*. And that suggests that in the SA-testing context, belief-expression is not adequate evidence for trusting the subject.

Indeed, the response from the Haidt subject at (3) *was* an expression of a first-order belief. It was this feature that, initially, made the response seem irrelevant to antagonism about OTJ* self-attributions. We decided that this was a superficial problem, however, since the utterance could plausibly be seen as shorthand for such a thing. Even so, in its original form, her expression would give parallel reasons for attributing her the first-order belief expressed. The explanation of the first-order action-expression would be the same type of explanation as before:

> (10) The subject uttered 'It [= the sex act] could result in emotional damage and birth defects', as opposed to something else, because:

(i) She believes that the siblings' act might have those results; plus
(ii) she desires to express her belief, and
(iii) she believes that her utterance is a way to express that belief.

And accepting the explanation means accepting that she has the first-order belief, as per (10.i). But in the SA-testing context, her first order belief clearly remains in doubt.

So in an SA-testing context, abduction from an apparently 'direct' expression of a belief is *per se* insufficient. In such a context, belief-expressions are inadequate evidence for trusting that the subject has the belief expressed. And this contradicts the antagonist's strategy for trusting the subject on her second-order beliefs.

As indicated, however, I do not think the violation of (NP) is a genuine problem. (Although again, if it is a real problem, that is no better for the antagonist.) After all, belief in the subject's second-order self-attributions still allows us to doubt meaningfully the first-order self-attributions. This is odd in one way, since there is no obvious reason why the 'order' of the judgment should matter to its reliability. Indeed, the evidence-based antagonist may see our favoritism toward higher-order beliefs as *ad hoc*. That may be why antagonism—and in particular, the corollary at (NP)—did not allow such a double-standard. (But see the next subsection for additional relevant dialectics.)

5.3 So What?

Jay Rosenberg used to joke that there were only two objections in philosophy. The first is 'No way' and the second is 'So what?' In the latter case, an objector will grant one's thesis, but deny that it has any real significance. In the present context, the antagonist may be similarly unimpressed about the experimenters' violation of (NP).

In particular, the antagonist can argue that the subject's self-understanding must be flawed in *some* substantive way regardless. After all, if the experimenters falsely presume that the subject gets her second-order beliefs right, the falsity of these presumptions *just means* the subject misapprehends some of her own beliefs (albeit second-order beliefs). On the other hand, if the experimenters' presumptions are correct, then per usual, the evidence suggests that the subject misattributes what she first-order believes. *So either way*, antagonism toward OTJ* self-attributions would be vindicated. A clever move by the antagonist.

The point, as one reviewer put it succinctly, is that if the antagonist is not entitled to rely on the subject's utterances, that already signals victory for fallibilism. So apparently, the antagonist is unharmed even if (NP) was *de facto* violated during the experiments.

Still, this seems like the antagonist is abandoning (NP). But more generously, perhaps she is just tweaking (NP) to acknowledge one exception. This modified (NP) would say that in SA-testing contexts, there can be 'no presumptions' that favor subjects' self-attributions, except when such presumptions serve only to vindicate antagonism in the end. The modified (NP) would still mean that the experimenters cannot presume subjects have the self-ascribed *first*-order beliefs. (In that case, antagonism would not be vindicated if such presumptions were true.) Yet it would mean experimenters can presume that the subjects have the second-order beliefs they express. Again, regardless of whether *that* kind of presumption is correct, antagonism would be warranted, as per the antagoinist's 'clever argument' above.

In reply, I grant that the antagonist's clever argument proves something uncongenial to infallibilism. It demonstrates that at least *some* OTJ* self-attributions are false. Even so, it does not follow that it is possible for any OTJ* self-attribution, whatsoever, to be false. And for that reason, it also does not follow that the evidence *shows* the fallibility of other OTJ* self-attributions. In particular, the antagonist's clever argument does nothing to discredit that experimental subjects are infallible about their occurrent, *second*-order beliefs, as reflected in their verbal expressions. Again, the experiments themselves assumed that a subject's grip on those beliefs was accurate.

One might even go a bit further. The experiments could be interpreted as even *supporting* the high reliability of (parts of) the subject's self-understanding. For the psychological studies are cases where intelligent researchers unearthed evidence against her self-conception, presuming all the while that she successfully expresses her own second-order judgments. And they presumed this *as a matter of course*. They not only trusted the subject here, they did so without independent evidence, and without flagging it as part of their methodology. Researchers apparently believed *it goes without saying* that we are justified in taking second-order belief-expressions as indicating the subject's actual second-order beliefs.

That antagonistically-inclined specialists adopted this trusting stance is a striking fact, one which should be explained. Perhaps, despite their credentials, they were making a methodological error. But naturally, another explanation is that our researchers *are* justified in trusting the subject to express her genuine second-order beliefs. Conceivably, their trust in the subject owes to a habituated linguistic practice, a practice that works precisely because speakers are (at least) highly reliable in this respect. Naturally, such an explanatory suggestion is contentious. But it is worth remarking that, oddly, the suggestion might gain support from our researchers' behavior, despite their antagonistic intent.

At the least, this illustrates that the exception-clause proposed by the clever argument is enough to relinquish 'evidence-based antagonism' as defined here. Again, given the background presumptions, the experiments do not debunk infallibilism about all OTJ*-attributions. To the contrary, the infallibilist hypothesis might even help rationalize the trust given to subjects.

More broadly, if any exceptions are made to (NP), that suffices to undercut the general claim at (A). For suppose it is sometimes rational (in an SA-testing context) to trust the subject's grip on her second-order judgments. Then, a test that trusts the subject in this way will not discredit every infallibility hypothesis. After all: Either the test will discredit the very trust it employs—or not. If so, then the test is self-undermining. In which case, it could not demonstrate that any infallibilist suggestion is false (though, per the clever argument, it may well show that some OTJ* self-attributions are false). On the other hand, suppose the test does not discredit the very trust it employs. Then, it is doubtful that the test discredits any infallibilist suggestion. For if it is rationally permissible for some OTJ* self-attributions to be presumed, and the test itself does not undercut that trust, nothing about the test tells against the infallibility of those self-attributions. (Besides, for practical purposes, the test treats those self-attributions as if they were infallible.) Thus, the test in this case does not discredit any infallibilist hypotheses about OTJ* self-attributions.

And so, an evidence-based antagonist apparently cannot allow exceptions to (NP), i.e., allow that trust in the subject is sometimes rationally permissible in an SA-testing context. Which is to say, if (A) is to be upheld, (NP) must be accepted in its full generality. This, by the by, is the earlier promised argument showing that (A) entails (NP). (For a more rigorous reconstruction of this argument, see the Appendix.)

6 Conclusion

The falsity of (NP) can come as a surprise, since again, it is odd that an experimenter must simply trust the subject in some instances, in order to judge whether she is wrong in other instances. That is so, even though there is good reason to say the subject *is* wrong in other instances. Even so, I suspect that psychology can do no better than this, at least at the present time. I cannot fully argue the point here; but briefly, if we do not trust the subject *at all*, then no utterance can be interpreted as expressing a thought that she actually has—*a fortiori*, expressing a mistaken second-order thought that she has. Perhaps some future experiment might detect such thoughts without depending on the subject's own verbal expressions. Perhaps neuroscience will advance to the point where we can just 'read off' the subjects' second-order thoughts from an fMRI. But the evidence from psychology is *de facto* not gleaned in these ways. As things presently stand, psychologists presume the subject is right in her self-understanding, re: the second-order judgments she expresses as if her own. And the rational permissibility of this, I claim, shows that the evidence fails to discredit every hypothesis suggesting some amount of infallibility.

However, it has likely occurred to the reader that there is a darker direction all this might take. Possibly, the unflattering results of the SA-tests suggest that it is *not* rationally permissible for psychologists to trust subjects' self-understanding. If so, then much of psychology would be bankrupt—soliciting verbal expressions from subjects would be rationally self-effacing. Moreover, it is unclear what the prospects for psychology would be, if such soliciting is debarred from its practice. In general, then, perhaps the real lesson should be that there is a *crisis of methodology* in experimental psychology.

In the foregoing, I have simply assumed that psychological experimentation is well-founded. And if a skeptic were to challenge this at this stage, I would have no quick rejoinder. One strategy, however, would be to argue for the infallibility of the subject's self-understanding, as reflected in some of her verbal expressions. If such a view were correct, then experimenters would not fall into error in trusting the subject's self-understanding. But to repeat, no positive argument for such infallibilism has been attempted here. The aim has just been to show that current psychology has not discredited such a view.

Appendix: Argument for the Entailment

We shall show that (A) entails (NP) by arguing the contrapositive: If~(NP), then~(A). But in lieu of 'the psychological evidence', I shall here talk of 'the results from *T*'

where *T* is the set of the tests from Nisbett & Wilson, Haidt, and the like (see n. 1). The argument is then as follows:

1. (NP) is false: There is a class *P* of self-attributions, whose members are presumed in at least one experimental test $t \in T$—and these presumptions are not irrational in their respective testing-contexts, despite the results from *T*. [Assume for conditional proof]

2. Results from *T* are evidence against some member of *P*, or not. [From 1, LEM]

3. If the results from *T* are not evidence against some member of *P*, then the results of *T* do not show that members of *P* are fallible. [Premise]

4. If the results from *T* are evidence against some member of *P*, then *T* is self-incriminating. [By the current usage of 'self-incriminating']

5. If *T* is self-incriminating, then the results from *T* do not show that all OTJ* self-attributions are fallible. [Premise]

6. So either way, the results from *T* do not show that all OTJ* self-attributions are fallible, i.e., (A) is false. [From 2–5]

Remarks:

Premise 3 is based on the thought that there must be evidence from *T* against a self-attribution, if its infallibility is to be discredited by the evidence. So, if *T* yields no evidence against members of *P*, the infallibility of these members is not discredited by the evidence.[23]

Regarding Premise 5, the background assumption is that self-incriminating experiments are limited in what they can show. After all, the experiments themselves indicate that something is faulty about the experiments. In which case, it is unlikely that they could show the fallibility of *every* OTJ* self-attribution.

Thus, if (NP) is false, then by the above proof, (A) is also false. In addition, (NP) is false, or so I have argued. Therefore, (A) is false. This, in a brief summary, is the case advanced in this paper.[24]

References

Anscombe, E. 1963. *Intention*, 2nd ed. Cambridge: Harvard University Press.

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

Bartlett, G. ms.: Occurrent states, unpublished draft.

[23] Some have objected that the experimenters are presuming the truth only of a subject's *token* self-attribution. And that does not conflict with (A), since it disparages hypotheses about a self-attribution *type*. But at 3, I have collected together the tokens into the set *P*, to suggest that the experiments do not debunk any infallibility hypothesis about tokens of the *P*-type. Granted, it has not been shown that *P* is a *natural kind*. But there is no reason to insist that the infallible self-attributions, if any, must form a *single* natural kind. There may be multiple kinds of self-attributions that turn out to be infallible.

[24] My thanks to Louise Antony, Dan Linford, William Lycan, Aidan McGlynn, and two anonymous referees for their valuable feedback on this material.

Bem, D. 1967. Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review* 74: 183–200.

Bortollotti, L., R. Cox, M. Broome, and M. Mameli. 2012. Rationality and self-knowledge in delusions and confabulations: implications for autonomy as self-governance. In *Autonomy and mental illness*, ed. L. Radoilska, 100–122. Oxford: Oxford UP.

Bortolotti, L. 2010. *Delusions and other irrational beliefs*. Oxford: Oxford University Press.

Bortolotti, L., and R. Cox. 2009. Faultless' ignorance: strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition* 18(4): 952–965.

Briñol, P., and R. Petty. 2003. Overt head movements and persuasion: a self-validation analysis. *Journal of Personality and Social Psychology* 84(6): 1123–1139.

Carruthers, P. 2011. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press.

Carruthers, P. 2014. On central cognition. *Philosophical Studies* 170: 143–162.

Chalmers, D. 2003. The content and epistemology of phenomenal belief. In *Consciousness: new philosophical perspectives*, ed. Q. Smith and A. Jokic, 220–272. Oxford: Oxford University Press.

Davidson, D. 1963. Actions, reasons, and causes. *Journal of Philosophy* 60(23): 685–700.

Davidson, D. 1983. A coherence theory of truth and knowledge. In *Kant oder Hegel?* ed. D. Henrich. Stuttgart: Klett-Cotta.

Dennett, D. 1975. True believers: the intentional strategy and why it works. In *Scientific explanations: papers based on Herbert spencer lectures given in the University of Oxford*, ed. A.F. Heath, 53–75. Oxford: Oxford University Press.

Devine, P.G. 1989. Stereotypes and prejudice: their automatic and controlled components. *Journal of Personality and Social Psychology* 56: 5–18.

Dutton, D., and A. Aron. 1974. Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology* 30: 510–517.

Epley, N., and T. Gilovich. 2005. When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making* 18: 199–212.

Festinger, L. 1957. *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Garrett, J., and C. Brooks. 1987. Effect of ballot color, sex of candidate, and sex of college students of voting age on their voting behavior. *Psychological Reports* 60: 39–44.

Gazzaniga, M. 1985. *The social brain: discovering the networks of the mind*. New York: BasicBooks.

Gazzaniga, M. 1995. Consciousness and the cerebral hemispheres. In *The cognitive neurosciences*, ed. M. Gazzaniga, 1391–1400. Cambridge: MIT Press.

Gopnik, A. 1993. How we can know our minds: the illusion of first person knowledge of intentionality. *Brain and Behavioral Science* 16: 1–14.

Grice, H. P. 1975. Logic and conversation. In *Syntax and semantics 3: speech acts*, eds. P. Cole & L. Morgan, 41–58. New York: Academic Press.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* 108: 814–834.

Hall, L., P. Johansson, B. Tärning, S. Sikström, and T. Deutgen. 2010. Magic at the marketplace: choice blindness for the taste of jam and the smell of tea. *Cognition* 117: 54–61.

Hall, L., T. Strandberg, P. Pärnamets, A. Lind, B. Tärning, and P. Johansson. 2013. How the polls can be both spot on and dead wrong: using choice blindness to shift political attitudes and voter intentions. *PLoS ONE* 8(4), e60554.

Horgan, T., and U. Kriegel. 2007. Phenomenal epistemology: what is consciousness that we may know it so well? *Philosophical Issues* 17: 123–144.

Johansson, P., L. Hall, S. Sikström, and A. Olsson. 2005. Failure to detect mismatches between intention and outcome in a simple decision task. *Science* 310(5745): 116–119.

Johansson, P., L. Hall, S. Sikström, B. Tärning, and A. Lind. 2006. How something can be said about telling more than we can know: on choice blindness and introspection. *Consciousness and Cognition* 15(4): 673–692.

Johansson, P., L. Hall, and S. Sikstrom. 2008. From change blindness to choice blindness. *Psychologia* 51(2): 142–155.

Kornblith, H. 2002. *Knowledge and its place in nature*. Oxford: Clarendon Press.

Kornblith, H. 2012. *On reflection*. Oxford: Oxford University Press.

Kornblith, H. 2013. Naturalism versus the first-person perspective. *Proceedings and Addresses of the American Philosophical Association* 87: 122–142.

Langland-Hassan, P. 2014. Unwitting self-awareness? *Philosophy and Phenomenological Research* 89(3): 719–726.

Lucas, E., and L. Ball. 2005. Think-aloud protocols and the selection task: evidence for relevance effects and rationalization processes. *Thinking and Reasoning* 11: 35–66.
Lycan, W. 1996. *Consciousness and experience*. Cambridge: MIT Press.
Merikle, P.M. 1992. Perception without awareness: critical issues. *American Psychologist* 47: 792–795.
Nisbett, R., and L. Ross. 1980. *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice-Hall.
Nisbett, R., and T. Wilson. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review* 8: 231–259.
Parent, T. 2007. Infalliblism about self-knowledge. *Philosophical Studies* 133(3): 411–424.
Parent, T. 2013a. Infallibility naturalized: reply to Hoffmann. *dialectica* 67(3): 353–358.
Parent, T. 2013b. In the mental fiction, mental fictionalism is fictitious. *The Monist* 96(4): 608–624.
Parent, T. Ms: Infallibilism about self-knowledge II: Paratactic judging. Available at http://www.unc.edu/~tparent/InfallibilismII.pdf.
Pronin, E. 2008. How we see ourselves and how we see others. *Science* 320: 1177–1180.
Pronin, E., D.Y. Lin, and L. Ross. 2002. The bias blindspot: perceptions of bias in self and others. *Personality and Social Psychology Bulletin* 28: 369–381.
Proust, J. 2013. *The philosophy of metacognition: mental agency and self-awareness*. Oxford: Oxford University Press.
Rosenthal, D. 2005. *Consciousness and mind*. Oxford: Clarendon.
Rubinoff, M., and D. Marsh. 1980. Candidates and color: an investigation. *Perceptual and Motor Skills* 50: 868–870.
Schwitzgebel, E. 2008. The unreliability of naïve introspection. *Philosophical Review* 117: 245–273.
Schwitzgebel, E. 2011. *Perplexities of consciousness*. Cambridge: MIT Press.
Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65(19): 555–567.
Smith, E., and F. Miller. 1978. Limits on perception of cognitive processes: a reply to Nisbett & Wilson. *Psychological Review* 85(4): 355–362.
Tversky, A. 1996. On the reality of cognitive illusions. *Psychological Review* 103(3): 582–591.
Tversky, A., and D. Kahneman. 1974. Judgement under uncertainty: heuristics and biases. *Sciences* 185: 1124–1131.
Vimal, R. 2009. Meanings attributed to the word 'consciousness': an overview. *Journal of Consciousness Studies* 16: 9–27.
Wason, P., and J. Evans. 1975. Dual processing in reasoning. *Cognition* 3: 141–154.
Wegner, D., and T. Wheatley. 1999. Apparent mental causation: sources of the experience of the will. *American Psychologist* 54: 480–491.
White, P. 1980. Limitations on verbal reports of internal events: a refutation of Nisbett & Wilson and of Bem. *Psychological Review* 87(1): 105–112.
Wilson, T. 2002. *Strangers to ourselves*. Cambridge: Harvard University Press.
Wright, C. 1998. Self-knowledge: the Wittgensteinian legacy. In *Knowing our own minds*, ed. C. Wright, B. Smith, and C. Macdonald, 13–46. Oxford: Oxford UP.
Wu, W. 2014. Being in the workspace, from a neutral point of view: comments on peter Carruthers' 'On central cognition'. *Philosophical Studies* 170: 163–174.