

Confabulation and rational obligations for self-knowledge

Sophie Keeling

To cite this article: Sophie Keeling (2018): Confabulation and rational obligations for self-knowledge, *Philosophical Psychology*, DOI: [10.1080/09515089.2018.1484086](https://doi.org/10.1080/09515089.2018.1484086)

To link to this article: <https://doi.org/10.1080/09515089.2018.1484086>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 26 Jun 2018.



Submit your article to this journal [↗](#)




Article views: 169



View Crossmark data [↗](#)

Confabulation and rational obligations for self-knowledge

Sophie Keeling 

Department of Philosophy, University of Southampton, Southampton, UK

ABSTRACT

This paper argues that confabulation is motivated by the desire to have fulfilled a rational obligation to knowledgeably explain our attitudes by reference to motivating reasons. This account better explains confabulation than alternatives. My conclusion impacts two discussions. Primarily, it tells us something about confabulation – how it is brought about, which engenders lively debate in and of itself. A further upshot concerns self-knowledge. Contrary to popular assumption, confabulation cases give us reason to think we have distinctive access to why we have our attitudes.

ARTICLE HISTORY

Received 3 October 2017
Accepted 2 March 2018

KEYWORDS

Confabulation; philosophy of mind; rationality; reasons; self-deception; self-knowledge

1. Introduction

When asked to explain why we have a given attitude, we make up answers in a striking number of cases – that is, we are prone to confabulate. I will argue that confabulation is motivated by the desire to have fulfilled a rational obligation to knowledgeably explain our attitudes by reference to motivating reasons. This account better explains confabulation than alternatives. My conclusion impacts two discussions. Primarily, it tells us something about confabulation – how it is brought about – which engenders lively debate in and of itself. A further upshot concerns self-knowledge. Contrary to popular assumption, confabulation cases give us reason to think we have distinctive access to why we have our attitudes.

This paper proceeds as follows: I first introduce confabulation and set out three explananda (Section 2) before highlighting gaps in existing explanations (Section 3). I then outline my own explanation (Section 4), before providing its benefits, which include its ability to address the explananda (Section 5). Finally, with my conclusion in hand, I emphasize its significance for understanding self-knowledge (Section 6).

CONTACT Sophie Keeling  sk5e14@soton.ac.uk  Department of Philosophy, University of Southampton, Avenue Campus Highfield Road, Southampton, Hampshire SO17 1BF, UK

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Confabulation

Confabulation cases are interesting, not least because they are often thought to help show that we lack *privileged access* to why we have our attitudes, that is, that we lack a distinctive and more reliable way of coming to know this about ourselves compared to other people (e.g., Nisbett & Wilson, 1977). Even those who think we have privileged access to some mental facts doubt these include the causes of our attitudes.¹ Confabulation cases might lead to this conclusion in several ways. One concerns the brute fact that self-ascription is less reliable than we might think. Another rests on how we explain such cases – we might think they are best explained by taking self- and other-knowledge to be fundamentally the same. This paper casts light on the second issue. I will return to the question of self-knowledge at the end, having argued for a particular account of confabulation. Along the way, I will also have said more about the Nisbett and Wilson-style picture of self-knowledge (Section 3.1). For now, though, let me keep my sights firmly on confabulation. In this section, I outline the phenomenon and set out some explananda that our explanation should address.

The paradigm example of confabulation is Nisbett and Wilson's stockings experiment; I will give more throughout the course of the paper.² Nisbett and Wilson arranged four pairs of identical stockings on a table and asked individuals which one they preferred. The majority picked those placed toward the right and so were influenced by what we can term the "position effect." At any rate, the subjects in general would not have formed their preference on the basis of a (perceived) reason – after all, the stockings were all the same. Yet, when asked why they preferred the pair they chose, the participants did not say that it was because of the position, or for no reason at all. Instead, they offered reasons like its "knit, sheerness, and weave" (Nisbett & Wilson, 1977).³ In such cases, the subjects fail to know why it is they have their attitudes – in this instance, they were ignorant that their preferences resulted from the position effect. But further, we can say that in providing the mistaken self-ascription, the subjects also *confabulate*.⁴

How best to define confabulation is contested, but here I am just interested in the mechanism underpinning (one particular subtype of) it.⁵ I shall, then, simply draw on aspects of Hirstein's (2005) definition and hope it, and the examples I will discuss, suffice to illustrate what I have in mind. Roughly, we might think that subjects confabulate in expressing an "ill-grounded" belief.⁶

A particular subtype of confabulation interests me here – the sort exemplified by the stockings experiment. First, it is an instance of *provoked* confabulation. The participants form the mistaken belief, and express it to the listener, specifically once they have been asked why they have the attitude. We can contrast this with the *spontaneous* sort in which subjects confabulate of their own volition. Further, the paper focusses on confabulation in non-clinical subjects,

rather than confabulation resulting from neurophysiological disorder.⁷ And finally, I will be concerned with confabulation about why we have our attitudes, which I take to be a substantial subtype – much of the literature surrounding non-clinical confabulation concerns the confabulation of explanations.⁸ I have, then, this in mind when referring to confabulation in what follows.

I now want to note three plausible explananda that an account of confabulation should address. I will then go on to show that my explanation, unlike others, helps us with all three.

Explanandum 1. Confabulators tend to mistakenly self-ascribe putative motivating reasons in particular, that is, they believe there is a reason for which they formed the attitude. We can contrast these with *purely causal explanatory* reasons which just explain the attitude by reference to causal factors such as the subject's character traits and biases.

While not always noted in the literature, this pattern is evident in the stockings experiment, and also in a range of other experiments.⁹

For example, subjects make this sort of mistake in at least some choice blindness studies.¹⁰ Choice blindness occurs when individuals select something, say an object or a theoretical position, and fail to notice when it is switched for another. In some experiments, the subjects are then asked why they picked the item (which, unbeknownst to them, they had not in fact selected). This then leads them to confabulate a motivating reason.

Take, for instance, Hall, Johansson, and Strandberg (2012). The experimenters told participants to mark their agreement with various ethical statements on a scale of 1 (*completely disagree*) to 9 (*completely agree*). Afterward, some of the statements were reversed and read to the subjects under the guise of what they had agreed or disagreed with. Subjects were frequently unaware of this swap, either at the time or when asked afterward, and “69% of all the participants accepted at least one of the two altered statement/rating relations” (p. 4). More relevant for us is what happened when the experimenters asked the subjects why they held this view (which they did not originally select). The individuals provided explanations which were false insofar as they pertained to a view that they did not hold.¹¹ These explanations took a particular form, and involved mistakenly attributing motivating reasons to themselves. For example, two participants originally agreed that “even if an action might harm the innocent, it can still be morally permissible to perform it.” The rating they gave was then reversed, e.g., from 9 to 1. The subjects tried to explain why they supposedly held the opposite position with the following:

No, no one should have to get hurt

No, well, I don't think it's ever ok ... I'm not exactly sure how to explain this, but innocents should never be hurt, you know, one should always find other ways of doing it. (Hall et al., 2012, supporting information, p. 1)

The participants, then, provided motivating reasons as opposed to a purely causal explanation. They did not, for example, reply with “I think hurting innocents is wrong because my parents drummed it into me and I’ve internalized the lesson well.”

Explanandum 2. Prompted confabulation sometimes occurs when no one is listening to the response.

Wilson, Dunn, Kraft, and Lisle (1989) performed various studies in which they asked subjects why they had certain attitudes. The subjects still confabulated (to themselves) even though the experimenters made clear they were not paying attention to the answer. The authors write that “We tell subjects in our studies that we want them to think about their reasons in order to organize their thoughts, and we explain that no one will ever read what they write” (p. 297).

Explanandum 3. Not only do we tend to mistakenly ascribe motivating reasons to ourselves, but we do so more readily than to others.

The type of confabulation that interests me here specifically concerns explanations of our own attitudes. The propensity to mistakenly ascribe motivating reasons seems to occur regarding self-ascription in particular. This is not to say whether we make more mistakes about our minds, but simply that we can see a pattern in the mistakes we do make which contrasts with other-ascriptions (both veridical and false). Unfortunately, Explanandum 3 has not been directly tested for. Nevertheless, I take the claim to be intuitive, and further, it receives support from the following two sets of studies.

First, we tend to provide motivating reasons when explaining our actions while observers give more purely causal explanatory reasons (Malle, Knobe, & Nelson, 2007).¹² We can see one instance of this pattern in Study 5 of Malle et al. (2007). Here, individuals were asked to “describe ‘the last time [they] had an interesting conflict with a romantic partner, friend, or parent’” (p. 502), and to explain a range of their and their opponent’s behaviors. Another participant, unrelated to the first, was also requested to explain the same behaviors based on the first subject’s account of the conflict. The actors’ explanations contained a greater number of motivating reasons compared to the observers’, and the observers’ explanations included a larger quantity of purely causal explanatory reasons. Indeed, this was the case regardless of whether the observer knew the actor (p. 503).

Admittedly, in investigating the explanation of action, the studies do not examine this paper’s specific concern – our *erroneous* explanations of our *attitudes*. Yet, on the basis of the experiment below, this pattern would likely occur in our explanations of attitudes as well.¹³ And given our tendency to confabulate, we can suppose that at least some of the self-ascriptions in the studies would have been false.¹⁴

Second, we tend to mistakenly think that we, but not others, form attitudes in a bias-free way. This stems from the bias blind spot, whereby we notice our own biases less than other people's (Pronin, Lin, & Ross, 2002).¹⁵ We can see this in the following study by Pronin et al. (2002). In it, pairs of subjects completed a "social intelligence test" with one being told their mark was above average, the other, below average. When asked to what extent they thought it was a good test and that its results would match up with those of similar ones, those who were told they did relatively well were likelier to appraise it higher on both fronts. Further, the subjects were then informed about "a self-protective tendency" that leads to such results in one's views about an assessment. Yet when asked, the individuals were more likely to take the other participant's views on the test as having been affected by their results, than to realize that the same could be said about their own.¹⁶ Individuals are on occasion, then, more likely to falsely maintain that their own judgments in particular do not result from purely causal explanatory reasons that indicate bias.¹⁷ Further, following earlier observations, we can suppose that if the subjects were asked why they made the judgment they did, they would confabulate motivating reasons in their own case but provide (correct) purely causal explanations of others' judgments.

We are more likely, then, to mistakenly use motivating reasons when explaining our attitudes compared to those of others. At the very least, it is a plausible prediction, and making it would be a mark in favor of an explanation of confabulation.¹⁸ All three explananda, I will argue in [Section 5](#), can be accounted for by my explanation, which sees confabulation as motivated by the desire to have fulfilled a rational obligation.

3. Alternative explanations of confabulation

Here I note the limitations of two prominent explanations before introducing my own in [Section 4](#). At any rate, it will be worth having some other options on the table so as to highlight the advantages of mine.

3.1. Mistaken inferential self-ascription account

One option is to accept inferentialism about self-knowledge of motivating reasons (or even inferentialism about self-knowledge more generally). We see this in the work of Cassam (2014), Carruthers (2013), Nisbett and Wilson (1977), Wilson and Dunn (2004), and Wilson (2002), for example. Indeed, such thinkers use confabulation cases as the data for an inference to the best explanation for this view. They argue that we confabulate in a given instance because we form our self-ascription using the same kind of inference that we use for other-ascriptions.¹⁹ Further, we also form knowledgeable self-ascriptions in

this way.²⁰ The only notable difference between self- and other-knowledge is that we have some additional evidence in our own case. This includes mental images, theories about how people normally form attitudes, thoughts, and feelings.

According to this view, we will confabulate if we process the evidence incorrectly or the evidence itself supports a false ascription. To return to our example, Nisbett and Wilson write that subjects in the stockings experiment explain their preference in terms of the pair's (supposed) sheerness because they use a misleading theory. The subjects assume that sheerness is a "representative reason" for preferring stockings, although it is not in fact why they prefer them in this case (1977, p. 249). *Sometimes* our mistakes might be motivated – in such instances, motivational factors would shape our inferences. Yet these would not carry out the main explanatory work, and at any rate, self-ignorance and error are generally unmotivated, such as in confabulation cases.²¹ It is important to the inferentialists' project that motivational explanations have limited applicability since, they take it, other accounts of self-knowledge could explain self-ignorance and error by appeal to them.²²

This view, though, is not best suited to account for Explanandum 3. If the same inferences underpin both self- and other-ascriptions, this raises the question of why confabulations follow a certain pattern we do not see in other-cases. The account's proponents, though, might make the following suggestion. We mistakenly self-ascribe motivating reasons when we would not ascribe them (correctly or incorrectly) to others because of the additional evidence we have concerning ourselves, such as mental images and feelings. Wilson, for example, writes that our extra evidence can sometimes serve as red-herrings, so to speak, and result in mistakes (2002, p. 108–110).²³

Here, though, we can say two things. First, Pronin and Kugler (2007) performed an experiment which suggests additional evidence does not cause the bias blind spot. It was similar to Pronin and colleagues (2002) as outlined above. This time, though, the experimenters gave some participants reports of what the other subject was thinking about when appraising the test. Yet, access to this information barely affected the degree to which individuals thought the other's score in the test influenced their judgment about it (2007, pp. 571–572).

Second, and more generally, we do not just have additional evidence suggesting that we form our attitudes on the basis of reasons. Some of it favors believing that we lack a reason for our attitude. For example, say the subject in the stockings experiment holds that people often prefer stockings on the basis of their sheerness. Yet, they have evidence that places doubt on the applicability of that theory in their own case – that they do not remember deliberating about the stockings before picking their favorite, say, or the fact that the stockings currently look the same to them.

Wilson and company, then, need to say why only some additional evidence influences our ascriptions.

At the very least, then, the *mistaken inferential self-ascription* approach needs to say more to account for Explanandum 3.

3.2. Current motivational accounts

Alternatively, we could think that motivational factors explain confabulation in one or both of two ways.²⁴ They might lead us to confabulate in the first place and/or cause our confabulation to have the particular content it does (see Sullivan-Bissett, 2015, p. 552) on this distinction).²⁵

I will go on to argue for a motivational account, but there is a limitation with one influential formulation. It concerns a motivational factor of the first sort: we confabulate because we are motivated by “simply the desire to avoid saying, ‘I don’t know,’ especially when the provoking question touches on something people are normally expected to know” (Hirstein, 2005, p. 17). Doing so would be “socially rewarded” (Bortolotti & Cox, 2009, p. 961) and would avoid “embarrassment” (Sullivan-Bissett, 2015, p. 555).²⁶ This, though, fails to explain Explanandum 2 – that individuals confabulate even when they think no one is paying attention to their answer.²⁷

Prominent options, then, fail to meet all our explananda. If there was an account that did, and did so easily and in a non-ad hoc way, then we would have a strong reason to accept it.

4. Confabulation and rational obligations

I argue that we confabulate because we are motivated by the desire to have fulfilled a certain rational obligation (we should not confuse this with moral obligations). In this section, I firstly set out the obligation I have in mind (Section 4.1), before outlining my explanation (Section 4.2). I will later say why we should accept it (Section 5), and discuss an upshot for self-knowledge (Section 6).

4.1. A rational obligation for self-knowledge

We should explain confabulation by appealing to what I call the *knowledgeable reasons explanation* obligation, or the *KRE* obligation for short:

Knowledgeable reasons explanation (KRE) obligation: The obligation to knowledgeably self-ascribe motivating reasons when explaining one’s own attitude.

We find this, and views in its vicinity, in Anscombe (2000), Boyle (2011a; section 4, 2011b), and Moran (2001, esp. pp. 124–129).²⁸

To give an example, say that I believe it will rain tomorrow. I ought to explain this in terms of my motivating reasons, and not purely causal explanatory ones. So, I ought to explain my belief by reference to, say, a motivating reason that the weather-person says it will rain. I ought not explain it in terms of my trust of authority figures, even if this explanation is valid. Further, this self-ascription – that I believe it will rain for the reason that the weather-person says it will – ought to be knowledgeable. That is, it ought to be true and not just a lucky guess. More can be said about the structure and grounds of this obligation, but I lack the space to discuss it here.²⁹

4.2. *The proposal*

We can use the obligation in the following proposal:

We confabulate, and indeed confabulate with the content we do, because we desire to have fulfilled the *KRE* obligation (i.e. the obligation to knowledgeably explain our attitudes by reference to motivating reasons).³⁰

In this subsection, I will briefly outline the rough picture before considering a more precise version.

According to my proposal, we confabulate when we lack an accessible explanation that would enable us to fulfill the obligation, that is, when we lack a motivating reason. I should note that this explanation is compatible with inferentialism about self-knowledge. The inferentialists could say that we confabulate because our desire to have fulfilled the obligation shapes our inferences. Yet my explanation of confabulation still differs from theirs. They see the cases we have been considering as non-motivational. Indeed, for the inferentialists, motivational factors do not perform the explanatory work even when they are present.

Let us consider an example and return to the panty-hose experiment. First things first, this seems to be the sort of situation in which subjects bear the undefeated *KRE* obligation, at least from their perspectives.³¹ We can further take it that the individuals desire to have met this obligation (I discuss both these moves in [Section 5.2](#)). The desire to have fulfilled the obligation leads the subjects to confabulate answers in the absence of a true one they can provide – they did not form them on the basis of reasons. And further, they specifically self-ascribe the reason that the stockings were sheerer, say, because it is a plausible motivating reason.

Now with the rough picture in hand, we can flesh it out with a possible mechanism. I want to be relatively open, but a good option would be to see confabulation as an instance of self-deception, and specifically self-deception construed along Alfred Mele's (2001) lines. This approach is appealing, as it requires few additional commitments. We already have independent reason to think that self-deception takes place. And further, Mele's account of it is

particularly economical since a lot of the work is performed by the operation of cognitive bias. The existence of bias is uncontroversial, and indeed we have already accepted it in this paper.

Mele offers the following account. Our desires can motivate self-deception since they lead us to underestimate and overestimate the importance of given pieces of evidence, pay more notice to certain pieces of evidence at the expense of others, and use particular methods of acquiring evidence, all in accordance with what would speak in favor of the result we want (pp. 26–27). And desires have this effect, Mele thinks, by interacting with cognitive biases (pp. 28–31). He mentions three such biases, the first concerning the “vividness of information.” When forming a belief, we are more likely to take information into account if it is vivid. Desiring something to be the case makes relevant pieces of evidence more vivid and can influence the resulting belief in this way. Second, we follow the “availability heuristic.” The ease with which we can recall tokens of a particular type (i.e., their availability) leads us to think they are disproportionately representative of that type. Since our motivations lead certain pieces of data to be more vivid, and vivid data is more available, our desires can influence belief formation by way of this heuristic as well. And third, the “confirmation bias” means that:

People testing a hypothesis tend to search (in memory and the world) more often for confirming than for disconfirming instances and to recognise the former more readily ... even when the hypothesis is only tentative (as opposed, e.g., to a belief one has). (p. 29)

And one’s desires influence the hypotheses one has, and therefore go on to confirm, since “favourable hypotheses are more pleasant to contemplate than unfavourable ones and tend to come more readily to mind”. (p. 30)

If we see confabulation as an instance of self-deception construed along Mele’s lines, we could understand the mechanism underpinning it more precisely, in the following way. Unbeknownst to ourselves, we mishandle the available evidence in line with our desire that we have fulfilled the *KRE* obligation. These desires lead subjects to both confabulate an answer to the questioner as opposed to admitting their ignorance, and to confabulate the specific content they do. Firstly, we might think that the subjects’ desire to have met the obligation leads them to overvalue the evidence, if there is any, that they formed their attitude on the basis of a reason. This may involve placing weight on:

- What they take to be a plausible normative reason. In using this piece of evidence, the subject would rely on the theory that: if we take x to be a normative reason for preferring p , x is our motivating reason for preferring p .
- The fact that one’s attitude can be based on a given reason without resulting from explicit deliberation.

And the subjects' desires may also cause them to undervalue or even ignore the evidence that they did not form their preferences on the basis of reasons. This might include the following facts:

- That they experienced uncertainty when considering potential normative reasons.
- That they cannot remember forming their preferences on the basis of the reasons in question, or even considering said reasons at all.
- That other people with the same attitude lack motivating reasons. When coupled with the claim that our minds work in similar ways to other people's, this might suggest that we lack them as well. For example, the subjects in the Pronin and colleagues (2002) study from Explanandum 3 may well have ignored this evidence, since their self- and other-attributions clearly differ.

In mishandling the evidence in this way, then, a subjects' desire to have fulfilled the obligation would lead them to adopt the relevant self-deceptive belief which they then express to the questioner.

Before I go on to motivate this explanation in [Section 5](#), I should end this section by clarifying why I appeal to desire to have fulfilled the *KRE* obligation, rather than several less committal alternatives. First, my explanation uses the *KRE* obligation as opposed to the related but less controversial obligation to form our attitudes on the basis of what we take to be normative reasons. This reference to responsible attitude formation at the lower-order level is the sort of thing that Pronin, Gilovich, and Ross, for example, seem to have in mind when discussing the possibility that the "biased [cognitive] searches" we engage in due to the bias blind spot "may blind us to our shortcomings and enhance our sense of rationality in a way that is undeniably ego enhancing" (2004, p. 788). We might simply say that subjects are motivated by the desire to have fulfilled the obligation to form their attitudes on the basis of motivating reasons. My obligation does presuppose this more minimal one. Yet, appealing to just the obligation to form attitudes on the basis of reasons only explains the content of the subject's confabulations. It does not explain why they confabulate in the first place – it is unclear why, under such a model, the subjects do not simply admit their ignorance. After all, to say that you do not know why you have an attitude it is not in itself to say anything about why you do actually have it. It may well be possible to have a motivating reason you are not aware of. Appealing to the *KRE* obligation, though, helps us explain both aspects of confabulation.

Second, my proposal states that we are motivated by the desire to have fulfilled an *obligation* to knowledgeably explain one's attitude with motivating reasons. An alternative explanation would simply be that we desire

to have knowledgeably ascribed motivating reasons without there being any normative demands to do so. But appealing to the *KRE* obligation provides a good, full, and simple explanation, and, as I will discuss in [Section 5.2](#), is independently plausible. On the other hand, if one explains confabulation simply using a desire to have knowledgeably ascribed motivating reasons, one still needs to say why subjects have this desire. Since it cannot be to impress others (recall [Section 5.2](#)), perhaps we might say that self-knowledge is some sort of “epistemic desideratum.” But this is not to say why a particular type of self-knowledge would be desired – why self-knowledge of motivating reasons as opposed to purely causal explanatory ones. Also, it is not obvious how we would cash out a desire for self-knowledge in non-normative terms. Perhaps we might say that self-knowledge is valuable to us because of pragmatic considerations. It helps us assess our attitudes and come to better decisions. And yet, knowing the truth – that we lack motivating reasons – would also be useful. Why, then, would subjects overlook the signs that they lack the relevant reasons in confabulation cases? It starts to look, then, that even if we could make the relevant maneuvers, they might be on the baroque side, and involve sacrificing simplicity.³²

5. Why accept this proposal?

Here I argue that the proposal explains confabulation particularly well. It has two main explanatory virtues: accounting for all the explananda ([Section 5.1](#)), and doing so in a non-ad hoc way with independently plausible components ([Section 5.2](#)).

5.1. Explanatory power

Unlike the alternatives we encountered in [Section 3](#), my explanation easily addresses all the explananda. First, it accounts for Explanandum 1 – that individuals self-ascribe specifically motivating reasons when asked why they have an attitude. After all, the proposal states that we desire to have fulfilled the *KRE* obligation, where this concerns motivating reasons. Second, if we recall, Explanandum 2 claimed that we confabulate even when lacking an audience. Yet, according to the proposal, it is not that we want to provide reasons so others think well of us. Rather, we want to have fulfilled the obligation to explain our attitudes by reference to motivating reasons. This is the case even if we are just explaining them to ourselves.

And third, Explanandum 3 stated that we confabulate in this way about ourselves but not others. Again, this is something my proposal explains. We desire to have fulfilled the obligation to knowledgeably explain *our own* attitudes by reference to motivating reasons, not other people’s.

Perhaps, though, one might object that the asymmetry is more naturally explained in terms of a broader pattern whereby we accurately, as well as mistakenly, explain our attitudes using motivating reasons; it is not an instance of self-deception at all.³³ After all, we saw in Malle and colleagues (2007) that subjects tend in general to provide more motivating reasons when explaining their own actions. But, insofar as individuals overlook evidence that they lack the ascribed motivating reason, their confabulations would seem to have a motivational element. And indeed, the subject has access to important clues – for example, that the stockings look the same, that they do not remember deliberating, that judgments can easily be influenced by our self-interest, and so on.

5.2. Plausibility

Further, the proposal accounts for the explananda in a non-ad-hoc way. As I will argue, that we believe we bear this obligation, and that we desire to have fulfilled it, are both already independently plausible (sections 5.2.1 and 5.2.2, respectively).

5.2.1. Our belief concerning the KRE obligation

First, our interpersonal interactions indeed suggest that we believe we bear the KRE obligation.³⁴ For example, say that I recently came to believe that Managua is the capital of Nicaragua and you ask me why. You would expect me to knowledgeably offer my grounds for the belief, for example, that a reliable website says so. You would see me as open to criticism if I replied with one of many alternatives. These include: “I don’t know,” “no reason,” “a reliable website says it’s the capital of Nicaragua, although that’s not *my* reason,” “a reliable website says so, but I’m probably wrong,” and “the perceptual mechanism detected patterns on a computer screen and processed them so as to result in a state of belief.”³⁵ Or, to consider a different attitude, say that I tried two yogurts and preferred the branded one to the supermarket offering. Again, you would think that something was wrong with my preference if I answered the question “why?” by saying something like “the advertising made it look like the sort of thing sophisticated people eat, and I want to be sophisticated.” Even if this were indeed true, you would still expect me to talk about the (supposed) rich and sophisticated flavors, and so on.³⁶

Here one might reply with various counter examples. Perhaps “no reason” is an appropriate response when enough time has passed that one might assume the subject has forgotten his or her original reason. Perhaps also it is acceptable when asked why I prefer chocolate cheesecake to strawberry, desire a Bakewell tart, or intend to stir my tea

counterclockwise, say. Yet, it is unacceptable in enough instances to suggest that the obligation is simply defeated in such counterexamples. They will be defeated, for instance, when the subject can assume that the listener already knows his or her motivating reason. The listener may already gather that attitudes formed a long time ago will often be based on the memory of evidence, which in itself is still a motivating reason. Insofar as “I don’t know” in these cases is an acceptable reply, it is because speakers just mean to communicate that they cannot remember what the initial evidence was. The obligation will also be defeated when the attitude is one that is not normally reason-sensitive, like a craving.³⁷ I cannot enumerate all the defeaters here, but for our purposes we can just note that the confabulation cases follow the same mold as the yogurt example, in which the obligation intuitively remains undefeated. As such, it is plausible to think that individuals will believe that they bear the *providing-* and *knowing-reasons* obligation in these instances.

Let me end this subsection by trying to assuage several worries we might have about attributing this sort of belief to the population at large. I should firstly note that I do not mean to over-intellectualize matters by saying that subjects believe they bear the *KRE* obligation. After all, the average person probably has not thought about these issues. Rather, I just want to suggest we have some sort of standing state that we can most simply capture in terms of the obligation. Indeed, introducing talk of the “*KRE* obligation” need not be ad hoc – take how we discuss moral reasoning. If subjects faced with the trolley problem say they would kill one person to save five, we might say they believe that one ought to bring about the greatest happiness for the greatest number. Indeed, based on more responses, we may even want to attribute a very fine-grained utilitarian principle to them, for example, concerning interests or preferences. But this is not to say that the subject thinks in those terms, or is in a position to explicitly provide such a principle. Second, there may also be individuals who do not even believe they are subject to the obligation in some undemanding *de re* sense. For example, certain philosophers will deny that we bear the obligation. Yet, I need not say that everyone has this belief, just that individuals will confabulate to the extent in which they do. There is space here for empirical research.

5.2.2. *Our desire concerning the KRE obligation*

Second, given that we believe we are subject to the obligation, it is also plausible we would desire to have fulfilled it. I can firstly note that we need not commit to anything very demanding regarding the desire in question. The desire could be as minimal as a tendency. Additionally, it is plausible that we would bear such a state regarding the *KRE* obligation and that it would play the role I am arguing it does. After all, doing as we ought to (generally) reflects well on us and it is reasonably uncontroversial to think

that we have a general tendency to see ourselves in a favorable light. For example, Wilson puts the general point as follows:

People's judgements and interpretations are often guided by ... the desire to view the world in the way that gives them the most pleasure – what can be called the “feel-good” criterion ... Just as we possess a potent physical immune system that protects us from threats to our physical well-being, so do we possess a potent psychological immune system that protects us from threats to our psychological wellbeing. When it comes to maintaining a sense of well-being, each of us is the ultimate spin doctor. (2002, p. 38)³⁸

We can draw on a range of data when saying that this wish to feel good manifests in positive self-appraisals. This includes some of the empirical support for self-deception, to the extent that it concerns self-deception about ourselves (e.g., Mele, 2001, pp. 3, 11). And we can also refer to various cognitive biases that have a similar effect. For example, the study concerning the bias blind spot that I outlined in [Section 3](#) showed our blindness to the consequences of the “self-serving bias,” in which people chalk their achievements down to themselves, but failures to other influences (Pronin et al., 2002, pp. 370, 377).³⁹ Indeed, it is a sign that something has gone wrong with the subject if they fail to view themselves and their circumstances through slightly rose-tinted spectacles, as we see with so-called “depressive realism.”⁴⁰

6. Upshots

We have good reason, then, to think that the desire to have fulfilled the *KRE* obligation motivates confabulation. This conclusion speaks to two debates. First, it tells us something about confabulation – that it is caused by a specific motivational factor. And second, it also has implications for understanding self-knowledge. That subjects systematically believe they bear the *KRE* obligation gives us at least some reason to think that we do bear the obligation, and further to accept agentialism about self-knowledge. My explanation also might help us to make a bolder claim, which I briefly mention at the end.⁴¹

That individuals systematically see themselves as subject to the *KRE* obligation is evidence that we do actually bear it. Indeed, taking the belief to be veridical is the most plausible and charitable option because it avoids attributing widespread error to individuals. It is also the simplest, since we do not then have to explain why subjects would all have made such a mistake.

If this is the case, it gives a reason to accept agentialism about self-knowledge. Agentialism can be understood in opposition to the inferentialist account of self-knowledge we encountered in [Section 3.1](#). To recall, the thought is that we learn why we have our attitudes (and other facts) by inferring it from various pieces of evidence about ourselves. The

inferentialists motivate their view by inference to the best explanation – confabulation occurs because the inferences underpinning our self-ascriptions can go wrong in various ways. Self-knowledge under this picture fundamentally resembles other-knowledge – the method is the same (inference), and there is nothing else significant to differentiate it. Agentialists, on the other hand, argue that this approach fails to capture the position of rational agency we bear in relation to our minds. As Gertler writes about the view (without endorsing it): “Some of our mental life expresses our agency – e.g., believing and intending are things we *do*. Moreover, recognizing an attitude as one’s own involves seeing it as a commitment for which one is responsible” (Gertler, 2015).

There are different formulations, but one version defended by Moran (2001) and Boyle (2011a, 2011b, 2009), might be captured by the following example⁴²:

Say that I ask you what you believe the capital of Nicaragua is. Considering what *you believe* the capital to be involves thinking about what it actually *is*. You might go on the internet, for example, or ask a knowledgeable friend. You will then realize that Managua is Nicaragua’s capital and can therefore reply that you believe that Managua is the capital city of Nicaragua. Further, you will also be able to say *why* you believe this – because both the internet and a reliable individual said as much.

We can acquire self-knowledge of our attitudes by using the *transparency procedure*. This is to answer the question of whether we believe that *p*, say, by answering the question of whether *p* is the case and thereby forming our belief on the matter (see esp. Moran, 2001, section 2.6). Further, in forming our belief that *p* on the basis of reasons in this way, we also come to know why we have it – on the basis of this consideration.⁴³

Importantly, part of this view is that, if we are unable to gain non-inferential self-knowledge of certain facts about ourselves, then we have done something wrong qua rational agents. It is not just that we use a distinct method for acquiring self-knowledge. For example:

We do not only allow [a subject’s self-ascription] to stand without the benefit of evidence, we also sometimes expect and sometimes insist that he take himself to be in a position to *speak for* his feelings and convictions . . . This normative expectation [“to speak for” your beliefs], and its relation to the rationality of the beliefs in question, certainly lends some support to the suggestion that the first-person accessibility of beliefs is not a *merely* empirical matter, an extra capacity for awareness of a certain class of facts we happen to have and whose absence would leave the psychological facts in question unaffected. (Moran, 2001, p. 26)⁴⁴

According to this picture, then, we bear rational obligations to be in a position to acquire non-inferential self-knowledge, and to do so using the transparency procedure. In this paper, I have provided evidence that we bear an obligation along these sorts of lines – the obligation to knowledgeably

explain our attitudes by reference to motivating reasons. Thus, we have (more) reason to accept agentialism.⁴⁵ This is the case even if we think that other accounts of self-knowledge can accommodate rational obligations for self-knowledge, which we might even doubt.⁴⁶ Still, these alternatives do not directly predict that we will bear something like the *KRE* obligation, unlike agentialism. Here I should acknowledge that confabulation suggests that self-knowledge may not be more reliable than other-knowledge. Perhaps we might have to depart from the traditional privileged access thesis I mentioned at the start. But I would be happy with that conclusion. Self-knowledge need not be more reliable than other-knowledge for the two to fundamentally differ. And greater reliability would be the least important part of privileged access under agentialism in any case.

Note that, so far, I have just said that my account of confabulation gives us at least some reason to accept agentialism. I have not yet said whether or not such cases give us *pro tanto* reason to accept the inferentialist position as well. More tentatively, I think they do not and that my explanation is compatible with accounts of self-knowledge other than inferentialism. A complication here, though, is that my preferred account of the mechanism for confabulation sees our desires influencing inferences underpinning the self-ascription. Yet, the proponent of privileged access can accept that self-knowledge is inferential, just at the subpersonal level. We can still appeal to a distinctive method for acquiring self-knowledge at the personal level of explanation.⁴⁷ In this way, self-knowledge would be like perception: underpinned by inferential low-level processing (e.g., Marr, 2010) and yet non-inferential at the level of the subject. This way of viewing self-knowledge, though, requires more work and Carruthers himself explicitly rejects this possibility (2013, p. 21–24). For the time being, then, I just want emphasize that at the very least confabulation cases do not unequivocally speak against privileged access, as has often been assumed.

7. Conclusion

This paper has argued that confabulation is motivated by the desire to have fulfilled the obligation to knowledgeably explain our attitudes by reference to motivating reasons. Accepting this provides us with a particularly satisfying explanation of confabulation, especially compared to alternatives. It is a rich conclusion, and bears significance for discussions of both confabulation and self-knowledge. In particular, it gives us at least some reason to think that we have distinctive access to our motivating reasons.

Notes

1. Cassam (2014) and Carruthers (2013) deny privileged access across the board, while Nisbett and Wilson (1977), Rey (2008), Gertler (2011), and Nichols and Stich (2003) are more circumspect.
2. The experimental reports are not always fully explicit on the details, but I am using standard interpretations here.
3. See also Wilson (2002, p. 103) and Wilson and Nisbett (1978, pp. 123–124).
4. One might deny that the subjects were mistaken about their reasons – perhaps the stockings actually seemed sheerer to them, and they preferred the pair on that basis. Sandis (2015), for example, responds to many such cases in this way. Yet, here I can say several things. First, this interpretation seems less charitable. It appears odd to think that subjects would take identical items to differ – perhaps they eventually come to prefer the stockings on the basis of their (perceived) sheerness, but it seems less plausible that they would do so from the start. Second, even if the (perceived) sheerness was their motivating reason, the subject was still mistaken in attaching so much explanatory importance to it. Nisbett later allows that the subjects in the stockings experiment might indeed have the self-ascribed motivating reason, but that nevertheless “by normal standards of discourse, [their] causal analysis is inadequate or incomplete” (Nisbett & Ross, 1980, pp. 217–219). And thirdly, we could bite the bullet in this case, but maintain that subjects still provide false self-ascriptions in others. At any rate, Sandis allows that other experiments show confabulation – ones which are especially important in the literature (e.g., the choice blindness cases I discuss later, and Haidt, 2001). Even if its scope is narrower than I hoped, my explanation will still hold in a significant type of case, then.
5. For overviews of how one might define confabulation, see Bortolotti and Cox (2009) and Hirstein (2009, 2005).
6. For example, we can contrast this with definitions of confabulation that just concern mistakes in memory; see Fotopoulou (2009) and McKay and Kinsbourne (2010).
7. For these distinctions, see Hirstein (2005).
8. Indeed, the confabulation of motivating reasons even constitutes Scaife’s (2014) definition. I should note that a reasonable amount of the literature on confabulated explanations concerns actions rather than our attitudes, see, for example, Hirstein’s (2009) section on confabulated introspection. This is something for further thought, but I expect my explanation regarding attitudes could be extended to these as well.
9. For examples of Explanandum 3, see Hall et al. (2012), Johansson, Hall, Sikström, and Olsson (2005), Johansson, Hall, Sikström, Tärning, and Lind (2006) (all choice blindness studies), and Haidt (2001). Also, for a similar pattern in explanations of action, see for example, studies involving split-brain patients (e.g., Gazzaniga, 2000) and hypnotized subjects (as discussed, e.g., in Wegner, 2002).
10. Thanks to Jordi Fernández and Ema Sullivan-Bissett for pointing out the applicability of these cases. See Scaife (2014, section 2.4) for a discussion of choice blindness in relation to confabulation.
11. I should note, though, that Hall and colleagues construe choice blindness in terms of the subjects’ attitudes actually changing (e.g., p. 5). Lopes (2014) also favors this interpretation. Yet, even if this is the case, the subjects still make a mistake about why they have this new attitude.
12. Malle and colleagues present the contrast in terms of “reasons” and “causal history” explanations, but I do not think our terminology differs substantively. On this asymmetry, see also Malle (2011) and Knobe and Malle (2002). This, they

- persuasively argue, is the best way of understanding the self and other asymmetry that some have tried to account for in terms of the *fundamental attribution error*, for example Jones and Nisbett (1972).
13. Jones (2002, pp. 227–228) and the cited study in Gilbert and Mulkey (1984) also suggest this. While the subjects are not explicitly explaining their own attitudes, when scientists were asked to explain others' errors, 'they all as if their own position is an unproblematic and unmediated re-presentation of the natural world. In contrast, the actions and judgments of those scientists who are depicted as being or as having been in error are characterized and explained in strongly contingent terms. Their false claims about the natural world are presented as being mediated through and as understandable in terms of various special attributes which they possess as individuals or as certain kinds of social actor' (Gilbert & Mulkey, 1984, p. 98).
 14. These results help address one worry we might have with the significance of Explanandum 1: perhaps confabulating subjects provide motivating reasons because the specific wording of the question "Why?" invites this response. As Sandis writes, when proposing this sort of concern with confabulation studies, experimenters are not always sensitive to this in the design and write up of their experiments (2015, pp. 270–271). For example, in the choice blindness study outlined earlier, Hall and colleagues write that subjects were "asked to explain *the reasoning* behind their ratings" (2012, p. 4, emphasis added). And in a similar study they report having "asked the participants to discuss and *justify* their ratings of the individual questions" (2013, p. 2, emphasis added). Yet in the same paragraph, they also write that subjects "were now asked to explain why" they made their choice (2013, p. 2), which is not at all leading. It is unclear whether they were, then, but if it turns out that subjects were explicitly asked for justification, the pattern in Explanandum 1 would be of little interest here. The subjects' self-ascriptions would have taken the shape they did because of the specific question asked. Yet, actors and observers in Malle and colleagues (2007) were presumably prompted using the same question but still gave different explanations. We can be confident, then, that there is a significant pattern that needs to be explained.
 15. Shermer (2012) also gives a nice summary.
 16. There was also a very slight tendency for subjects to be blind to the presence of what was labeled "a 'self-protective' tendency" in themselves but not others, but this was not statistically significant.
 17. Pronin does write elsewhere that "sometimes the 'bias blind spot' is primarily caused by people's unwarranted denials of their own biases, whereas at other times it is more attributable to people's overestimations of others" (Pronin, 2007). The subjects in the above study, though, do underestimate of the influence of bias in themselves (see Pronin et al., 2002, p. 377), and therefore do make false self-ascriptions.
 18. See also Nisbett and Wilson (1977, p. 273) for a case in which subjects recognize the possibility that others' tolerance for electric shocks might have been manipulated by the experimenter, but not their own.
 19. I use 'inferentialism' to refer to this sort of view, as opposed to accounts appealing to a distinctive kind of inference, such as Byrne (2011).
 20. Carruthers is most explicit in making this move from thinking confabulations are inferential to thinking all self-ascriptions are (2013, Ch. 11).
 21. Cassam writes that "your failure to perform the necessary inference or your misinterpretation of the evidence *might* be motivated but needn't be" (2014, p. 195).
 22. See Cassam (2014, pp. 193–194) and Carruthers (2013, pp. 337–338). They discuss this regarding self-ignorance of what our attitudes are.

23. Pronin and Kugler (2007, p. 566) consider this option as a way of explaining the bias blind spot, but reject it.
24. The approaches in Sections 3.1 and 3.2 map onto the distinction between seeing confabulation either as resulting from “dysfunction” in normal knowledge acquisition or from a “compensating mechanism” due to a failure to know (Fotopoulou, 2009; p. 246; see also McKay & Kinsbourne, 2010; p. 291; both discuss this in terms of the confabulation of memory, but we can also put it in more general and epistemic terms).
25. Indeed, they are often thought to play some sort of role, contra the views of those in Section 3.1. For example, Hirstein (2005), Bortolotti and Cox (2009, pp. 954–955), and Sullivan-Bissett (2015, p. 562) see motivational elements as an important, although not necessary, aspect of confabulation.
26. See also McKay and Kinsbourne (2010, p. 291) and Fotopoulou (2009, pp. 270–271) for discussions of this sort of view.
27. Carruthers uses the data from Explanandum 2 to make this point (2013, pp. 338–339).
28. It can be hard to determine what exactly these philosophers commit themselves to. This is especially the case since the *KRE* obligation resembles the (putative) obligation to justify our attitudes. This differs from *KRE*, though, since it would not require the subject to ascribe the relevant considerations as the reasons for which they formed the attitude. For example, it would be acceptable to answer the question “why?” with “*p*, although it is not *my* reason.” Also, the authors are not always explicit that the self-ascriptions ought to be knowledgeable. Boyle (2011a; section 4) comes the closest, although he allows that matters could also be cashed out in terms of a “positive epistemic status” other than knowledge (p. 8). Still, I can draw support from the general considerations given in these works.
29. For example, fulfilling the *KRE* obligation seems to require actually having motivating reasons for the attitude. In this way, it appears importantly connected to a (putative) obligation to form one’s attitudes on the basis of reasons.
30. I am flexible on the precise formulation of this general thought. We might instead think that we are motivated by the desire to *believe* that we have fulfilled the obligation. On this distinction in terms of self-deception, see Nelkin (2002) and Fernández (2013). Also, I prefer talking of the desire to *have* fulfilled it, as opposed to the desire to fulfill it, but again I am not committed to this.
31. Sometimes the obligation will be undefeated in confabulation cases, but sometimes it might not. My explanation only requires, though, that the circumstances in these instances are sufficiently like those in normal cases that subjects would plausibly believe they bear the obligation.
32. Nevertheless, while I think we should not, I would be happy if one accepts this explanation. I would still have argued for a motivational account of confabulation, and that the relevant motivation concerns self-knowledge. This in itself is significant. And further, it would still give rise to the upshot I discuss in Section 6.
33. See Cox (2018) for an account of this sort. Appiah (2009) also considers something like this as an explanation of self-ignorance, but dismisses it (p. 43).
34. In this argument, I am following Boyle (2011a; p. 236, 2011b; p. 10, 2009; pp. 4–5). In defense of this general sort of claim, see also Moran (2001) and, regarding action, Anscombe (2000).
35. This latter response might be appropriate in certain situations (e.g., scientific discussions). Still, automatically offering it in a normal context seems to express a peculiar relationship to your belief. Jones (2002) is relevant here.

36. It is worth noting that the motivating reasons we expect people to self-ascribe can be very minimal. In the case of perceptual belief, say, it might be enough simply to give replies shaped by Pryor's dogmatism or Huemer's phenomenal conservatism – one might self-ascribe the motivating reason that it “seems to [me] as if p is the case” (Pryor, 2000, p. 519) or “seems to [me] that p ” (Huemer, 2007, p. 30). My point is simply that, defeaters non-withstanding, we expect people to knowledgeably self-ascribe at least *some* motivating reasons.
37. On judgment-sensitive desires, see Scanlon (1998).
38. See also Gilbert and Wilson (2000).
39. See also Coleman (2015, “self-serving bias”), and Turner and Hewstone (2009). Other biases include: the *positivity bias*, *unrealistic optimism*, and the *Lake Wobegon effect*; see Coleman (2015).
40. For example, see Brown (2007), although see Moore and Fresco (2012) for caution in the precise details of the theory.
41. For another account of how confabulation cases might lend support to agentialism (under a different name), see Cox, 2018. I prefer my own strategy, though, for considerations mentioned at the end of Section 5.1.
42. Alternatively, see Burge (1999, 1996).
43. This account of the self-knowledge of motivating reasons comes from Boyle:

“If I reason ‘P, so Q’, this must normally put me in a position, not merely to know that I believe Q, but to know something about why I believe Q, namely, because I believe that P and that P shows that Q” (2011b, p. 8). I have used this for the agentialist picture of how we know our motivating reasons for simplicity's sake. That said, I prefer the following picture: you learn why you believe that p by considering what the reasons are for believing that p (as opposed to considering whether to believe that p). That is, the question of your reasons is transparent to the question “why believe that p ?” We can see something of this in Moran. (2001, p. 127)
44. It is less clear whether Moran also thinks that we are obligated to be in a position to know our motivating reasons, but it certainly fits nicely with his picture. If anything, though, my paper suggests that agentialists should pay more attention to *KRE* obligation.
45. The more minimal alternative raised at the end of Section 4.2 also helps support agentialism. Say we deny that subjects are motivated by the desire to have fulfilled an epistemic *obligation*, but instead simply by a desire for self-knowledge. Yet agentialism also predicts that we would desire to be able to acquire self-knowledge – because of the *KRE* obligation or something similar.
46. Gertler (2011, Ch. 8; 2016) argues that we need not accept a rational agency account to do this.
47. Another option might be to say that only confabulatory self-ascriptions are inferential and that self-knowledge is underpinned by just the distinctive method. Yet, as Carruthers argues (2013, Ch. 11), privileged access theorists struggle to explain why subjects would use inference as opposed to the distinctive method in confabulation cases. And my account of confabulation does not help us here. The following initially looks like an appealing move: subjects employ inference because they desire to have fulfilled the *KRE* obligation and the distinctive method would tell them otherwise. Yet, this requires the subject to already recognize at some level that they lack an accessible reason, which is implausible and unparsimonious.

Acknowledgments

I would like to thank the two anonymous reviewers from Philosophical Psychology, Alfonso Anaya, Ryan Cox, Richard Gray, Eleanor Gwynne, Conor McHugh, Ema Sullivan-Bissett, Lizzy Ventham, and Jonathan Way. I am also grateful to audiences in Cardiff, Southampton, the Self-deception: what it is and what it is worth conference in Basel 2017, the Philosophy of Language and Mind conference in Bochum 2017, the Transparency in Belief and Self-Knowledge workshop in Oviedo 2015, and the Annual Meeting of the European Society for Philosophy and Psychology in Tartu 2015.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Arts and Humanities Research Council [AH/L503939/1].

Notes on contributor

Sophie Keeling is a postgraduate researcher in philosophy at the University of Southampton, and is co-supervised by the University of Southampton and Cardiff University. She holds a BA in English Literature and Philosophy from Cardiff University and an MA in Philosophy and Literature from the University of Warwick. Her interests lie primarily in the philosophy of mind and epistemology (especially self-knowledge).

ORCID

Sophie Keeling  <http://orcid.org/0000-0002-7846-453X>

References

- Anscombe, G. E. M. (2000). *Intention* (2nd ed.). Cambridge, MA: Harvard University Press.
- Appiah, K. A. (2009). *Experiments in ethics*. Cambridge, MA: Harvard University Press
- Bortolotti, L., & Cox, R. E. (2009). 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, 18(4), 952–965.
- Boyle, M. (2009). Two kinds of self-knowledge. *Philosophy and Phenomenological Research*, 78(1), 133–164.
- Boyle, M. (2011a). 'Making up your mind' and the activity of reason. *Philosophers' Imprint*, 11(17), 1–24.
- Boyle, M. (2011b). Transparent self-knowledge. *Proceedings of the Aristotelian Society Supplementary Volume*, 85(1), 223–241.
- Brown, J. (2007). *The self*. New York: Psychology Press.
- Burge, T. (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society*, 96(1), 91–116.

- Burge, T. (1999). A century of deflation and a moment about self-knowledge. *Proceedings and Addresses of the American Philosophical Association*, 73(2), 25–46.
- Byrne, A. (2011). Transparency, belief, intention. *Proceedings of the Aristotelian Society Supplementary Volume*, 85(1), 201–221.
- Carruthers, P. (2013). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.
- Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.
- Coleman, A. M. (2015). *A dictionary of psychology* (4th ed.). New York: Oxford University Press.
- Cox, R. (2018). Knowing why. *Mind & Language*, 33(2), 177–197. doi10.1111/mila.12173.
- Fernández, J. (2013). *Transparent minds: A study of self-knowledge*. Oxford: Oxford University Press.
- Fotopoulou, A. (2009). Disentangling the motivational theories of confabulation. In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy* (pp. 263–290). New York: Oxford University Press.
- Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7), 1293–1326.
- Gertler, B. (2011). *Self-knowledge*. New York: Routledge.
- Gertler, B. (2015). Self-knowledge. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2015 Ed.). Retrieved from <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/>
- Gertler, B. (2016). Self-knowledge and rational agency: A defense of empiricism. *Philosophical and Phenomenological Research*, 96, 1–19.
- Gilbert, D. T., & Wilson, T. D. (2000). Miswanting: Some problems in the forecasting of future affective states. In J. P. Forgas (Ed.), *Thinking and feeling: The role of affect in social cognition* (pp. 178–197). Cambridge: Cambridge University Press.
- Gilbert, G., & Mulkay, M. (1984). *Opening Pandora's box*. Cambridge: Cambridge University Press.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834.
- Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE*, 7(9).
- Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (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).
- Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge, MA: MIT Press.
- Hirstein, W. (2009). Confabulation. In T. Bayne, A. Cleermans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 174–177). Oxford: Oxford University Press.
- Huemer, M. (2007). Compassionate phenomenal conservatism. *Philosophical and Phenomenological Research*, 74(1), 30–55.
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745), 116–119.
- Johansson, P., Hall, L., Sikström, S., Tärning, 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(4), 673–692.
- Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the cause of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). Morristown: General Learning Press.

- Jones, W. E. (2002). Explaining our own beliefs: Non-epistemic believing and doxastic instability. *Philosophical Studies*, 111(3), 217–249.
- Knobe, J., & Malle, B. F. (2002). Self and other in the explanation of behavior: 30 years later. *Psychologica Belgica*, 42(1–2), 113–130.
- Lopes, D. M. (2014). Feckless reason. In G. Currie, M. Kieran, A. Meskin, & J. Robson (Eds.), *Aesthetics and the sciences of mind* (pp. 21–36). New York: Oxford University Press.
- Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. *Advances in Experimental Social Psychology*, 44(1), 297–352.
- Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor–Observer asymmetries in explanations of behaviour: New answers to an old question. *Journal of Personality and Social Psychology*, 93(4), 491–514.
- Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.
- McKay, R., & Kinsbourne, M. (2010). Confabulation, delusion, and anosognosia: Motivational factors and false claims. *Cognitive Neuropsychiatry*, 15(1–3), 288–318.
- Mele, A. (2001). *Self-deception unmasked*. Princeton: Princeton University Press.
- Moore, M. T., & Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6), 495–509.
- Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton; NJ: Princeton University Press.
- Nelkin, D. K. (2002). Self-deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly*, 83(4), 384–406.
- Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. New York: Oxford University Press.
- Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice-Hall.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.
- Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, 11(1), 37–43.
- Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, 111(3), 781–799.
- Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, 43(4), 565–578.
- Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3), 369–381.
- Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34(4), 517–549.
- Rey, G. (2008). (Even higher-order) intentionality without consciousness. *Association Revue Internationale De Philosophie*, 1(243), 51–78.
- Sandis, C. (2015). Verbal reports and ‘real’ reasons: Confabulation and conflation. *Ethical Theory and Moral Practice*, 18(2), 267–280.
- Scaife, R. (2014). A problem for self-knowledge: The implications of taking confabulation seriously. *Acta Analytica*, 29(4), 469–485.
- Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Belknap.
- Shermer, M. (2012). *The believing brain: From ghosts and gods to politics and conspiracies – How we construct beliefs and reinforce them as truths*. New York: St Martin’s Press.
- Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition*, 33, 548–560.

- Turner, R. N., & Hewstone, M. (2009). Attribution biases. In J. M. Levine & M. A. Hogg (Eds.), *Encyclopedia of group processes & intergroup relations* (pp. 43–45). Thousand Oaks, CA: SAGE.
- Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
- Wilson, T. D. (2002). *Strangers to ourselves*. Cambridge, MA: Belknap.
- Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. *Advances in Experimental Social Psychology*, 22, 287–343.
- Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, 55(1), 493–518.
- Wilson, T. D., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology*, 41(2), 118–131.