

ARTICLE

 OPEN ACCESS

## The puzzle of transparency and how to solve it

Wolfgang Barz

Department of Philosophy, Goethe-University, Frankfurt am Main, Germany

### ABSTRACT

According to the transparency approach, achievement of self-knowledge is a two-stage process: first, the subject arrives at the judgment ‘*p*’; second, the subject proceeds to the judgment ‘I believe that *p*.’ The puzzle of transparency is to understand why the transition from the first to the second judgment is rationally permissible. After revisiting the debate between Byrne and Boyle on this matter, I present a novel solution according to which the transition is rationally permissible in virtue of a justifying argument that begins from a premise referring to the mental utterance that is emitted in the course of judging ‘*p*.’



**ARTICLE HISTORY** Received 11 October 2018; Accepted 2 January 2019

**KEYWORDS** Transparency approach to self-knowledge; justification; belief; Alex Byrne; Matthew Boyle

The puzzle of transparency is a by-product of the transparency approach to self-knowledge according to which one comes to know one’s own mental states, not by peering inward, but by focusing on the aspects of the external world that one is aware of in virtue of having the mental states in question. Roughly stated, the puzzle is this: how should the focus on external states of affairs (such as the location of Toronto) bring about knowledge about my mental states (such as my belief that Toronto is located in Ontario)? Naturally, this only puzzles people who think that the transparency approach is by and large correct. If you think that the transparency approach is wrong-headed anyway, then there is no puzzle for you. Nonetheless, the puzzle is of interest even to opponents of the transparency approach since, if there were no satisfying solution to the puzzle, then this would be a point against the transparency approach. Thus, even opponents of the transparency approach may find the following beneficial.

The notion of transparency is notoriously ambiguous: different philosophers associate different ideas with it. To forestall possible misunderstanding, then, it is essential to rule out those meanings of ‘transparency’ which are not relevant in this context. I conduct this task in [Section 1](#). Furthermore,

---

**CONTACT** Wolfgang Barz  [barz@em.uni-frankfurt.de](mailto:barz@em.uni-frankfurt.de)  Department of Philosophy, Goethe-University, Norbert-Wollheim-Platz 1, Frankfurt am Main 60629, Germany

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2019 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-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

I tentatively defend the transparency approach against a widespread objection. In [Section 2](#), I present the puzzle of transparency and explain what is so puzzling about it. In [Section 3](#), I outline a solution proposed by Byrne (2005) and revisit Boyle's (2011) criticism of it. Finally, in [Section 4](#), I present my solution to the puzzle which is, in a sense, a syncretic proposal that tries to reconcile the opposing views of Byrne and Boyle.

## 1. Some preliminary remarks on the notion of transparency

First, let me emphasize that by 'transparency' I will not refer to the traditional, sometimes called 'Cartesian,' claim that if a person is in a certain mental state, then this person is in a position to know that she is in that state. This is the idea that Ryle (1949) once dubbed 'self-intimation' and Williamson (2000) today calls 'luminosity.' However, self-intimation and luminosity have nothing to do with the idea of transparency, at least as used here. In this paper, the notion of transparency refers to the idea that, typically, a person comes to know that she is in a particular mental state, not by peering inward at her mental state itself, but by focusing instead on certain aspects of the external world. Let us call transparency in this sense – the sense that is relevant here – self-ascriptive transparency.

Self-ascriptive transparency should be distinguished from both phenomenal transparency and transparency of doxastic deliberation. Phenomenal transparency is the claim that, when I try to attend to the phenomenal properties of one of my perceptual experiences, I end up attending to some features of mind-independent objects. Phenomenal transparency is most prominently emphasized by Tye (1995). Tye says, for example, that, if I try to attend to the phenomenal color of my visual experience of a ripe tomato, I end up attending, not to some feature of my experience, but to the red surface of the ripe tomato. Notice that phenomenal transparency might be related to, but is far more specific than, self-ascriptive transparency – for phenomenal transparency is restricted to situations in which I try to discover the type and content of my own perceptual experiences. In contrast, self-ascriptive transparency is not so restricted but also covers situations in which I try to discover the type and content of other mental states, especially so-called propositional attitudes such as belief and desire.

However, I will not expand on all the different types of propositional attitudes but will focus on belief.<sup>1</sup> The claim of self-ascriptive transparency on which I will concentrate is that we come to know that we believe that  $p$  – where  $p$  stands for a proposition about the external world – not by peering inward and rummaging through our belief-box, so to speak, but by focusing instead on the external-world-proposition that serves as the content of our belief. In other words, the idea is that I come to know that I have a particular

belief, not by focusing on anything mental, but by focusing exclusively on the worldly items my belief is about.

Self-ascriptive transparency in this sense is related, but not quite the same as transparency of doxastic deliberation, as discussed by Shah and Velleman (2005). Whereas transparency of doxastic deliberation is the claim that I can answer the question whether I *should* believe that  $p$  by answering the question whether  $p$ , self-ascriptive transparency is the claim that I can answer the question whether I, *in fact*, believe that  $p$  by answering the question whether  $p$ . Thus, there is some similarity here, but the notion of self-ascriptive transparency is not the same as the notion of transparency of doxastic deliberation.

Having differentiated between different meanings of transparency, I will use the term 'transparency' in the following to cover only self-ascriptive transparency. The other types of transparency will play no role in what follows.

One well-known source of the transparency approach to self-knowledge (henceforth 'TA') is an often-cited passage from Evans's *Varieties of Reference* (1982) according to which I answer the question whether I believe that there will be a third world war by answering the question whether there will be a third world war. The essence is that, to achieve doxastic self-knowledge, as it may be called, I do not have to look inside but at the world. Thus, the usual and everyday method of knowing one's own beliefs – which has been traditionally considered a kind of *introspection* – turns out as a specific variety of *extrospection*.

To be sure, talk of 'extrospection' should not be taken too seriously here. In particular, proponents of TA are not committed to the view that we come to know our own beliefs by sense perception. Instead, TA only implies that, typically, we come to know our own beliefs by *some* outward-oriented mental activity that may, but need not be sense perception. According to Moran (2001), for example, the outward-oriented mental activity consists in *weighing the evidence* for and against a particular proposition. To use Evans' original example for illustration, suppose that I am asked whether I believe that there will be a third world war. According to Moran, I answer this question by weighing the evidence for and against the prospect of a third world war. Suppose, for example, that I may find that the reasons for the proposition *that there will be a third world war* outweigh the reasons against it. Consequently, I will make the judgment that, yes, there will be a third world war. This, in turn, drives me to the conclusion that I, in fact, believe that there will be a third world war. Call this the 'evidence-based account' of the outward-oriented mental activity in which we engage to determine our own beliefs.

The evidence-based account has some serious drawbacks.<sup>2</sup> The objection is that considering the reasons for and against the proposition  $p$  will, at least in many cases, create a new belief rather than uncover an already existing

one. Suppose, for example, that I never thought about the possibility of a third world war before the question of whether I believe that there will be a third world war was directed at me. Thus, at the time of the question, I did *not* believe that there will be a third world war. However, after deliberating on the issue, I find that the reasons for the prospect of a future world war outweigh the reasons against it. Consequently, I conclude that I *have* the belief in question, though I did not have that belief at the time the question was directed at me. Similarly, the procedure described by the evidence-based account may result in ignoring beliefs that I have. Just think of beliefs that are not in line with the evidence at my disposal, such as religious beliefs, deeply ingrained prejudices, or superstition. By weighing the evidence at one's disposal, one will, hopefully, realize that one should refrain from believing propositions not supported by one's evidence. Accordingly, one will conclude that one *lacks* the beliefs in question, though one actually has them – at least at the time before the process of deliberation started.

In light of these objections, it is often argued that TA is ill-conceived and has to be superseded either by an improved version of the inner sense theory (Gertler), by some kind of expressivism (Finkelstein) or even by a Ryle-inspired inferentialism (Cassam). However, this reaction is a bit overhasty. Even if Moran's evidence-based account fails, this does not imply that TA goes down the tube. In my view at least, there is a promising alternative to the evidence-based account that remains perfectly in accord with the basic idea of TA. According to this alternative, the outward-oriented mental activity by which we come to know our own beliefs does not consist in weighing the evidence for and against a particular proposition, but just in *understanding* the proposition in question. Nishi Shah and David Velleman nicely summarize the procedure as:

“posing the question whether *p* and seeing what one is spontaneously inclined to answer. In this procedure, the question whether *p* serves as a stimulus applied to oneself for the empirical purpose of eliciting a response ... [T]he procedure requires one to refrain from any reasoning as to whether *p*, since that reasoning might alter the state of mind that one is trying to assay. Hence, asking oneself whether *p* must be a brute stimulus in this case rather than an invitation to reasoning” (Shah and Velleman 2005, 506).

To use Evans's example for illustration again, suppose you are asked whether you believe that there will be a third world war. Furthermore, suppose that you believe that there will be a third world war. Now, recall that having a belief to the effect that *p* implies having the disposition to judge that *p* whenever the issue arises (and conditions are favorable). So, given that you believe that there will be a third world war, once you understand the proposition that there will be a third world war, your disposition to make the respective judgment is triggered and you spontaneously judge – that is, you judge without reflecting

on any evidence – that there will be a third world war. Consequently, you conclude that you believe that there will be a third world war. This account, which may be called ‘spontaneous judgment account,’<sup>3</sup> manages without the idea of deliberation or assessment of reasons or weighing evidence. Accordingly, it is immune to the objections that are commonly raised against Moran’s evidence-based account. So the failure of Moran’s evidence-based account does not engulf TA in the abyss. The spontaneous judgment account is still there as a fallback option.

Indeed, the spontaneous judgment account is not without difficulties either. One might wonder, for example, whether one’s spontaneous judgment about a certain matter is a reliable indicator of one’s belief.<sup>4</sup> However, I will not enlarge upon this question – for my overall aim in this paper is not to defend a specific variant of TA, but to discuss a puzzle, the puzzle of transparency, which pertains to both the evidence- and the spontaneous judgment account. My aim at this point was only to tentatively defend TA against a widespread but, in my view, unwarranted objection and thereby pave the way for exposing the puzzle of transparency, to which I will turn in the next section.

## 2. The puzzle of transparency

To get a better understanding of the puzzle, it may be helpful to abstract from the differences between the evidence-based and the spontaneous judgment account and to emphasize their similarities instead. According to both accounts, the achievement of doxastic self-knowledge is a two-stage process. At the first stage, the subject arrives at a judgment in response to a specific outward-oriented question. At the second stage, the subject proceeds from this judgment to a further judgment, this time about her own belief. Let us take a look at those two stages in slow motion. Suppose, for example, that a subject who believes that Toronto is located in Ontario wonders whether she believes that Toronto is located in Ontario. According to both the evidence-based and spontaneous judgment account, the subject can answer this question by answering the outward-directed question whether Toronto is located in Ontario. So, if things go well, then the subject will – be it due to a process of deliberation or not – arrive at the judgment that Toronto is located in Ontario.

However, making the judgment that Toronto is located in Ontario is not the end of the story – for this judgment does not state anything about the attitude the subject has towards the proposition *that Toronto is located in Ontario*. It simply says that Toronto is located in Ontario, which is just a fact about how things stand in the external world. It does not say anything about how things stand *with the subject*; it does not say that the subject *believes* that Toronto is located in Ontario. So, to arrive at a state of self-knowledge, the subject has to move on from the judgment that Toronto is

located in Ontario to the judgment that *she herself believes* that Toronto is located in Ontario. Let us call this latter judgment 'second-order judgment,' and the former judgment 'first-order judgment.'

Now, a new question arises: why is it *rational* for the subject to proceed from the first-order judgment that Toronto is located in Ontario to the second-order judgment that she believes that Toronto is located in Ontario? It does not seem obvious that the transition from a judgment about a geographical fact to a judgment about the mental state of a specific person is rationally permissible. On the contrary, it seems that the transition from 'Toronto is located in Ontario' to 'I believe that Toronto is located in Ontario' is utterly bad! To begin with, the proposition *that Toronto is located in Ontario* does not logically imply the proposition *that I believe that Toronto is located in Ontario*. There is not even an empirical supporting relation between the first and the second proposition. To empirically infer from the fact that Toronto is located in Ontario, that there is some specific person in Middle-Europe, namely Wolfgang Barz, who has a particular belief about Toronto's location seems quite bizarre. There is just no empirical connection, no law of nature or such that guarantees or even makes it probable that Wolfgang Barz has a certain belief, given Toronto's location. In other words: There is just no suitable relation between the two propositional contents – the content of the first-order and the content of the second-order judgment – that could explain why the transition from the first- to the second-order judgment is epistemically admissible, good, or something that is rational to do. However, proponents of TA are committed to the view that the transition is admissible, good, and rational. Hence, there is a problem.<sup>5</sup>

Note that the puzzle of transparency is not supposed to be a skeptical challenge to self-knowledge. The motivation behind the problem is not to suggest that it is impossible to know our own beliefs. Rather, the motivation behind the problem is to emphasize that proponents of TA owe us an answer to the question of why the transition from a judgment about the external world to a judgment about one's own mind is rationally permissible. Unless proponents of TA provide an answer to this question, their dialectical position is quite weak – for who wants to accept an account of self-knowledge that portrays us as relying on an irrational belief-forming method? Thus, the problem is not a general skeptical problem on par with, say, Boghossian's (1989) trilemma about self-knowledge. It is a specific problem for proponents of TA.

This becomes especially clear if it is realized that for proponents of inner sense models there is no puzzle at all. When I make the judgment that Toronto is located in Ontario, then, let us suppose, there is a token representation of the proposition *that Toronto is located in Ontario* somewhere in my mind. According to proponents of inner sense models, I can take notice of this representation: I can see it flashing up in my mind, so to speak. This, it could be said, is the reason why the transition from 'Toronto is located in Ontario' to 'I believe that Toronto is located in Ontario' is justified. However, proponents of TA cannot take this line.

They may agree that, when I make the judgment that Toronto is located in Ontario, then there is a representation of the proposition *that Toronto is located in Ontario* flashing up in my mind. However, according to proponents of TA, we must do without any special faculties of inner sense, self-scanning devices, or self-monitoring capacities. Thus, we lack the capacity to look inside and see which mental representations are flashing up there – mental representations are just hidden from our view. So, the puzzle of transparency might be described as the challenge of explaining why it is rational to proceed from a judgment about the external world to a judgment about one's mind, *given that we lack any capacity to look inside and directly observe our mental representations*.

Now, after having outlined the puzzle, I will turn to its solution. To set the stage for my proposal, I would like first to review a suggestion made by Alex Byrne and then consider Matthew Boyle's criticism of it.

### 3. Byrne's proposal and Boyle's criticism

At the heart of Byrne's (2005) proposal lies the observation that anyone who proceeds from the judgment that  $p$  to the judgment that she herself believes that  $p$  can be described as following an epistemic rule. An epistemic rule is a hypothetical imperative which states that, if some specific condition is satisfied, then you should do this or that. In the case of proceeding from the judgment that  $p$  to the judgment that I believe that  $p$  the respective epistemic rule reads:

(BEL) If  $p$ , judge that you believe that  $p$ !<sup>6</sup>

As Byrne points out, anyone who follows this rule necessarily ends up with a true judgment. To follow BEL the subject must first determine whether the condition mentioned in BEL's antecedent is satisfied. This is done by considering whether  $p$  is the case (this corresponds to the first stage of achieving doxastic self-knowledge that I distinguished above.) Now, if the subject comes to the conclusion that  $p$  is the case and that, thereby, the condition mentioned in the antecedent is satisfied, then she is in a state of believing that  $p$ . Thus, says Byrne, BEL is self-verifying: unless the subject believes that  $p$  she cannot follow the rule. Following the rule presupposes that the subject believes that  $p$ . So, following the rule guarantees that the subject will end up with a true second-order judgment.

Byrne is entirely correct regarding the self-verifying character of BEL. However, I doubt whether this provides an answer to the question 'Why is it rational for the subject to proceed from a first- to a second-order judgment?' The fact that BEL is self-verifying implies that proceeding from a first- to a second-order judgment is highly reliable. However, being *highly reliable* is not sufficient for being *rationaly permissible*. This is at least suggested by Laurence Bonjour's (1980) case of the reliable, but unknowing, clairvoyant, that is, someone who has the power to scry facts

reliably without knowing that he possesses this power. The mere fact that the clairvoyant's beliefs are reliably produced does not suffice to make his beliefs rationally permissible. From the clairvoyant's perspective, his beliefs come out of the blue. That his beliefs turn out to be true must seem a sheer accident from the clairvoyant's perspective. It seems that the situation of a subject who follows BEL is similar to the situation of the unknowing clairvoyant: the subject forms her second-order judgment by means of a highly reliable method, but – unless she is aware that the method is highly reliable – she finds herself at a loss how to make sense of the truth of her second-order judgment: as we saw earlier in this paper, there is just no suitable evidential relation between the propositional contents of first- and second-order judgments. Matthew Boyle nicely puts the point: 'a modicum of rational insight will inform me that, even if it is true that  $P$ , this by itself has no tendency to show that I believe it.' Thus, Boyle continues, 'Byrne's ... approach ... represents the subject as drawing a mad inference' (Boyle 2011, 230–231). So something is missing from Byrne's account: Byrne does not tell us, why, *from the perspective of the subject who engages in the transparency procedure*, to proceed from first-order to second-order judgments is a rational thing to do.

Perhaps one may be inclined to think that there is an easy way out for Byrne here: just add – besides the condition that the subject forms her second-order judgment in accordance with a self-verifying rule – the further condition that the subject also *realizes* that she acts in accordance with a self-verifying rule. It might be suggested that, once this further condition is met, the subject is in a position to see that proceeding from first- to second-order judgment is rational. Recall that this is the move that BonJour once recommended regarding the clairvoyant: to make the clairvoyant's beliefs rational from the clairvoyant's perspective, just provide him with the knowledge to the effect that his beliefs are formed using a highly reliable method.<sup>7</sup> However, this strategy will not work in the case at hand because to realize that one acts in accordance with BEL, one needs to know that one's second-order judgment is based on one's first-order judgment. This, in turn, implies that one knows that one judges, and hence believes, that  $p$ . Thus, the strategy already presupposes the kind of self-knowledge that needs to be explained in the first place.

In response to this dilemma, Matthew Boyle (2011) has suggested an alternative view, the so-called 'reflective approach,' which tries to explain the inherent rationality of the transition from first- to second-order judgment without depicting this transition as an inference. According to Boyle, believing  $p$  and knowing oneself to believe  $p$  are not two different mental states, but two aspects of the same mental state. When I believe something, says Boyle, then I am tacitly aware of the fact that I believe it. So, when I pass from first- to second-order judgment, I consciously acknowledge what



I already tacitly knew. In light of this account, the transition from first- to second-order judgment is not a kind of inference from one propositional content to another, but an act of reflection, an act of making explicit or articulate a piece of knowledge that there was all along. According to Boyle, this explains why proceeding from first- to second-order judgment is rational from the perspective of the subject.

There is something right and something wrong in Boyle's view. Boyle is right when he claims that Byrne represents the subject as drawing a mad inference and, thus, cannot explain why the transition from first- to second-order judgment is rational from the subject's point of view. However, I find it hard to accept Boyle's view that believing  $p$  and knowing oneself to believe  $p$  are aspects of one and the same mental state. I cannot believe that anytime we form a belief about the external world, we are tacitly aware of ourselves as believing the proposition in question. I have no argument, but I think that this persistent virtual self-awareness, as it may be called, is phenomenologically implausible. At least from my experience, there are many situations in which I form a belief about the external world absentmindedly, that is, without any, even tacit, awareness of myself as having the belief in question.<sup>8</sup>

However, I will not harp on about this point because, even if one accepts the view that we are always tacitly aware of our beliefs, it is still unclear how the process of making this tacit self-awareness explicit is supposed to work without violating the requirement to do without any inwardly directed monitoring or detecting capacities. As a proponent of TA, Boyle cannot say that what he calls 'reflection' proceeds by way of observing or taking notice of the content of our tacit self-awareness. Nevertheless, Boyle seems to endorse such a view:

"The reflective approach explains doxastic transparency ... as a matter ... of shifting one's attention from the world with which one is engaged to one's engagement with it – an engagement of which one was already tacitly cognizant" (2011, 228).

To my ears at least, this sounds pretty much like a rejection of TA. If doxastic self-knowledge is achieved by a shift of attention from the outer to the inner realm, then the idea of transparency gets lost. Thus, to retain the plausible aspects of Boyle's criticism of Byrne without giving up TA, we need to formulate a third alternative: an alternative that does not represent the subject as drawing a mad inference, yet avoiding notions incompatible with TA. This is precisely what I will attempt in the next section.

#### 4. A novel solution to the puzzle

First, I would like to introduce some auxiliary assumptions on which I will rely in what follows. The most important assumption is that making a judgment is the mental analog to making an assertion. Making an assertion contains at

least three elements: first, the asserter – the person who makes the assertion; second, the vehicle of assertion, which are the words uttered by the asserter; and third, the assertive force with which the words that the asserter utters are uttered. I assume that these three elements are also present in the case of judgments, especially that there is a mental analog to the vehicle of assertion: mental words and mental sentences that are mentally uttered or thought with assertoric force. It may seem as if the admission of mental words commits me to some demanding empirical hypothesis such as Fodor's language of thought. However, I do not think that I am thus committed because the assumption that there are mental words does not imply that those words are physical structures in the brain. They may be structures of a Cartesian mental substance instead. So the claim that there are mental words as such is neutral concerning the physicalism/dualism-debate in the philosophy of mind and hence should be acceptable for philosophers of all stripes. The picture I would like to suggest, then, is this: when someone makes a judgment to the effect that  $p$ , he or she thinks some mental sentence that means that  $p$  with assertoric force – where the verb 'to think' refers to the mental analog of the activity of uttering.<sup>9</sup>

Let us return to the question of why the transition from first- to second-order judgment is rational from the perspective of the subject. Suppose again that I make the judgment that Toronto is located in Ontario – either as the result of a process of deliberation or as the spontaneous outcome of understanding the proposition in question. According to the picture I just drew, I think a sentence that means that Toronto is located in Ontario with assertoric force. Now, my strategy to solve the puzzle of transparency is to formulate a justifying argument that is anchored in the mental sentence that is thought when making the judgment that Toronto is located in Ontario. The argument will only employ rather trivial premises that any normal subject can justifiably believe without engaging in inner observation. So, if the argument succeeds, it will explain why the transition from first- to second-order judgment is rational from the first-person perspective without violating the requirement to do without any inwardly directed monitoring or detecting capacities.

If, as I assume, making the judgment that Toronto is located in Ontario consists in mentally uttering a sentence that means that Toronto is located in Ontario, it must be possible for me to refer back to that mental utterance in subsequent thought, just as during a conversation with someone else it is possible for me to refer back to one of his or her utterances. Moreover, since there could be no doubt that I understand my own mental utterance, it must be possible for me to ascribe to it the property of *being a sincere assertion to the effect that Toronto is located in Ontario*. This will be the first premise of the justifying argument that I am about to formulate:

(1) **That** is a sincere assertion to the effect that Toronto is located in Ontario.

Notice that the word 'that' in bold letters does not refer to the proposition *that Toronto is located in Ontario*, but to the mental utterance of the words, 'Toronto is located in Ontario,' that, according to my assumption, takes place when making the judgment that Toronto is located in Ontario.

It might be objected that I could not refer back to an utterance – whether mental or not – unless I focused on its syntactic features such as the shape or sound of the words uttered. However, it should be clear that proponents of TA cannot admit this because it would violate the requirement to do without presupposing any inwardly directed monitoring or detecting capacities. So it seems that premise (1) is not admissible by the standards of TA.

My response is that I do not accept the view that one could not refer back to an utterance unless one focused on its syntactic features: it suffices that one attends to its *meaning*. Imagine that, immediately after I publicly uttered the words 'Toronto is located in Ontario' with assertoric force, you forget about the shape and sound of my words so that you cannot even tell whether I spoke English, German, French or whatever. As long as you grasped the meaning of my words, however, you will still be able to refer back to them – for example, by using a description such as 'the last utterance in our conversation that meant that Toronto is located in Ontario.' To refer back to my utterance, you do not need to focus on or remember its shape or sound; you only have to understand it. The same goes for mental utterances: you do not need to take notice of their syntactic features (whatever they are); you only have to understand them. The activity of understanding the meaning of one's own utterances – whether mental or not – in turn, does not presuppose any act of inner observation or any inwardly directed monitoring or detection. It only presupposes the mastering of language. Thus, premise (1) is admissible in the present context.

It might be objected that I am too quick to draw this conclusion. In order to understand a public utterance, it is necessary to notice it first. If you don't see or hear an utterance, you cannot understand it. Seeing or hearing an utterance amounts to grasping its syntactic features. Now, if mental utterances are analogous to public utterances, then there must be some mental analog to the activity of taking notice of an utterance, call it 'mental noticing.' Mental noticing, in turn, is a kind of inwardly directed monitoring capacity whereby subjects grasp the syntactic features of their current mental utterances. Thus, it seems that my account is doomed to fail from the outset: the postulation of mental utterances brings in its train the acceptance of an inwardly directed monitoring capacity.<sup>10</sup>

My reply is that it might be true that, in order to understand a *third person* public utterance, it is necessary to notice it first. However, it is not true that, in order to understand one of *my own* public utterances, it is

necessary to notice it first. In the first person case, understanding and uttering occur within the same act, at least as far as assertions are concerned: to make an assertion is to utter words you already understand; you don't need to notice your words before you can make sense of them. From the first person perspective, the sense is already there. Thus, the postulation of mental utterances is innocent: it does not commit us to the existence of an inwardly directed monitoring capacity.

However, premise (1) is still suspicious in another respect. Recall that premise (1) not only presupposes that I refer back to some utterance; it also presupposes that I ascribe the property of *being a sincere assertion* to it. So the question is: how am I to *know* that the mental utterance to which I refer back is a sincere assertion? It might be said that I do not know that an assertion made by another person is sincere unless I know that the person who makes the assertion believes what she says. Further, it might be said that the same goes for one's own assertions. Thus, it seems that knowing the sincerity of my own assertion requires me to know that I believe that Toronto is located in Ontario. That is to say, premise (1) already presupposes the truth of the claim that the justifying argument (I am about to formulate) is supposed to establish. Thus, it seems that premise (1) is not admissible for reasons of circularity.

However, you need not know that the speaker believes what she says to be justified in believing that she is sincere. It suffices that you are not in possession of evidence to the effect that the speaker is insincere. At least, this is the default rule in normal conversational situations: as long as you do not detect any signs of insincerity on the side of the person opposite, you are justified in believing that she is sincere – you do not need to establish first what the person believes. I see no reason why this rule, when it comes to one's own mental utterances, should cease to be in force. Recall that the hypothetical subject of my example believes that Toronto is located in Ontario. Since, in this case, the subject and the speaker are the same person, there are no possible signs of insincerity that the subject might detect. Consequently, the subject is justified in believing that the speaker – who happens to be herself – is sincere.<sup>11</sup> Thus, premise (1) is admissible. There is no danger of circularity here.

Up to this stage of the argument, the subject is licensed to conclude that there is *someone* who makes a sincere assertion to the effect that Toronto is located in Ontario. However, the subject is not licensed to conclude that *she* is the one.<sup>12</sup> Thus, a further premise is needed:

- (2) The person who utters **that** is identical to the person who utters *this*, and I am the person who utters *this*.

Again, the word ‘that’ in bold letters refers back to the original utterance of the mental sentence ‘Toronto is located in Ontario.’ In contrast, the word ‘this’ in italics – as it appears both in the first and the second conjunct – refers to premise (2) itself. Let us take a look at the first conjunct of (2) first. The first conjunct identifies the utterer of the original mental sentence, ‘Toronto is located in Ontario,’ with the utterer of (2). The question is: how does the subject know that the utterer of the original sentence is identical to the utterer of (2)? At first sight, it might be tempting to suppose that the subject cannot know whether two given utterances have their source in the same person unless the subject has taken a close look at both utterances and carefully compared their features, such as their characteristic sound or tone. However, this line of thought would lead into a blind alley again, because comparing characteristics of mental utterances would presuppose an ability to overhear those utterances and notice their characteristic features with an inner ear. Thus, the unwelcome commitment to inwardly directed monitoring capacities would return.

Fortunately, there is another way of justifying the first conjunct of premise (2) that does without any such unwelcome commitments. I take my cue here from a paper by Enoch and Schechter (2008) in which they address the question of why we are justified in employing basic belief-forming methods such as inference to the best explanation, modus ponens or reliance on perception. Enoch and Schechter argue that we are justified in employing these methods in virtue of the fact that employing them is indispensable for successfully engaging in activities that are central to rationality. Such activities include understanding the world around us, deliberating about what to do, planning for the future, and so on. If, for example, I would refrain from employing any inference to the best explanation, I would not be able to make sense of the external world. That implies that I could not engage in a rationally required project. Thus, the method of inference to the best explanation is essential for being rational. In a sense, then, basic belief-forming methods are justified in virtue of their pragmatic indispensability.

This idea naturally extends to *beliefs*. If a belief is such that holding it is a necessary precondition for being rational, then one is justified in holding it, even in the absence of any positive evidence. The first conjunct of premise (2) expresses such a belief: it attributes of two mental utterances, which are in fact one’s own, the property of having common authorship. Consider what happened if one would lack beliefs of this type, that is, if one would not believe of mental utterances, which are in fact one’s own, that they have common authorship. In my view, this would amount to a state of mental disorder similar to thought insertion, that is, a state in which the subject feels as if the thoughts that are hers stem from someone else. Being in a state of mental disorder similar to thought insertion is a severe obstacle to engaging in the activity of reasoning. However, not being able to engage

in the activity of reasoning would deprive us of the kernel of our rationality. Thus, believing of two mental utterances that come up in one's mind that they have common authorship is indispensable for being rational. The first conjunct of premise (2), then, is justified in virtue of its pragmatic indispensability. The subject does not need to gather any positive evidence.

At this point, a severe difficulty looms.<sup>13</sup> It might be said that the strategy I adopt to justify the first conjunct of premise (2) – the 'Enoch and Schechter strategy' for short – might be used to solve the puzzle of transparency in a simple and straightforward way. It seems plausible to say that transitions from first- to second-order judgments are indispensable for successfully engaging in activities that are central to rationality. If we would not carry out this transition, we would not be able to achieve doxastic self-knowledge. However, we need to achieve doxastic self-knowledge in order to understand ourselves. Thus, the transition in question is pragmatically indispensable for being rational, just as other fundamental belief-forming methods are, such as inference to the best explanation, use of modus ponens or reliance on perception. Now, if the Enoch and Schechter strategy is right when applied to the first conjunct of premise (2), it cannot be wrong when applied to transitions from first- to second-order judgments. So, why not choose the Enoch and Schechter strategy instead of constructing a justifying argument to solve the puzzle? Why not claim that the transition from first- to second-order judgment is justified in virtue of its pragmatic indispensability?

The problem with this proposal is that the Enoch and Schechter strategy amounts to a form of externalism. In order to be justified in the Enoch and Schechter way, employment of the belief-forming method in question has to be indispensable for being rational. However, the subject does not need to *believe* that employment of the belief-forming method is indispensable for being rational. According to Enoch and Schechter, a subject might be justified in employing a belief-forming method even if she does not have access to the fact due to which she is justified.<sup>14</sup> Hence, the Enoch and Schechter strategy is of no help for explaining why the transition from first- to second order judgments is rational *from the subject's point of view*.<sup>15</sup>

This response may provoke a second, more worrisome, concern: why should we think that the method for acquiring doxastic self-knowledge should be subject to a more rigorous standard of rationality than other basic belief-forming methods? If what Enoch and Schechter (2008) argue with respect to inference to the best explanation, use of modus ponens, reliance on perception, and other basic belief-forming methods is correct, and I am right that the Enoch and Schechter strategy is a form of externalism, then it seems that subjects are perfectly justified in using those methods without their being rational from the subject's perspective. Why shouldn't we conclude that the same goes for our method for acquiring doxastic self-knowledge? Why shouldn't we conclude that

we are perfectly justified in proceeding from first- to second-order judgments despite the fact that this transition is not rational from the subject's perspective? It seems, then, that I apply double standards: as regards inference to the best explanation, use of modus ponens, reliance on perception, and other basic belief-forming methods I seem to tolerate the view that a belief-forming method may be justifiably employed without being rational from the subject's perspective; however, when it comes to the method for acquiring doxastic self-knowledge I insist on the view that it cannot be justifiably employed unless it is rational from the subject's perspective. This seems to be inconsistent at least.

Let me say in response that, according to TA, the method for acquiring doxastic self-knowledge is not a *basic* belief-forming method on a par with inference to the best explanation, use of modus ponens and reliance on perception. Recall that, according to TA, doxastic self-knowledge does not require some special-purpose epistemic capacity (such as inner observation) that comes in addition to other epistemic capacities. Instead, the method for achieving doxastic self-knowledge is parasitic upon other belief-forming methods such as perception and inference. TA is, in Byrne's terminology, *economical* in that it explains doxastic self-knowledge 'solely in terms of epistemic capacities and abilities that are needed for knowledge of other subject matters.'<sup>16</sup> Thus, even if proponents of TA accept the Enoch and Schechter strategy in regard to basic belief-forming methods, they are not committed to accepting it in regard to the method for achieving self-knowledge – for this latter method is non-basic.

What about the second conjunct of premise (2): 'I am the person who utters *this*' – where the word '*this*' refers to the utterance in which the word '*this*' appears? The second conjunct of premise (2) is simply true by definition of the first-person pronoun 'I'. Any token of 'I' refers to the person who is the author of the utterance in which that token appears. Because it makes this definition explicit, the second conjunct of premise (2) is an *a priori* truth. Again, there is no need for postulating any inwardly directed monitoring or detecting capacities to explain how the subject can know (or be justified in believing) it.

The rest of the justifying argument falls rather easily from the assumptions already accepted. From (1) and (2) it follows that:

(3) I sincerely assert that Toronto is located in Ontario,

from which, backed by the trivial claim that whenever someone sincerely asserts that  $p$  she believes that  $p$ , it follows that I believe that Toronto is located in Ontario, QED.

Note that the justifying argument just outlined does not start from the proposition *that Toronto is located in Ontario*. Instead, it starts from a premise referring to a specific mental utterance. Thus, the proposition *that Toronto is located in Ontario* makes no direct contribution to the

argument. However, it makes an indirect contribution: it is vital for the argument that the mental utterance to which the first premise refers expresses the proposition *that Toronto is located in Ontario*. So, even if there is no suitable evidential relation between the proposition expressed by '*p*' and the proposition expressed by 'I believe that *p*,' this does not imply that there is no reasonable way from the judgment '*p*' to the judgment 'I believe that *p*.' The point of my proposal is that the act of judging '*p*' is accompanied by certain beliefs which evidentially support the proposition expressed by 'I believe that *p*.'

According to my proposal, then, the justification involved in the achievement of doxastic self-knowledge is perfectly normal inferential justification. Thus, it might be objected that we typically do not come to know our own beliefs by inference: our judgments about our own beliefs are *spontaneously* formed. However, this objection is based on a misunderstanding concerning the idea of inferential justification. The idea of inferential justification concerns the way a belief is *justified* – it does not concern the way a belief is *formed*. Note, especially, that to be inferentially justified in holding the belief that *p*, one does not need to *infer* that *p* from the premises of which the justifying argument consists. Rather one only needs to *justifiably believe those premises*. It is not even required that those beliefs are occurrent.<sup>17</sup> So, in a sense, inferential justification can go without inference. Therefore, it is not self-contradictory to claim that there are judgments which are both spontaneously formed and inferentially justified. Second-order judgments are a case in point.

Nonetheless, it might be objected that the conditions on doxastic self-knowledge that follow from my proposal are still too strong – for, even if the subject does not need to infer the conclusion 'I believe that Toronto is located in Ontario' by explicitly going through all steps of the justifying argument, the subject needs to *believe* premises (1) and (2) at least. And this might seem highly implausible. Aren't those premises just figments of the philosopher's imagination? I do not think they are. Premise (1), recall, gives expression to our capacity for understanding our own thoughts. Premise (2), in turn, gives expression to our capacity for reasoning, particularly the capacity for treating one's own thoughts as having their source from the same subject and thinking of that subject as oneself. So, even if the wording of premises (1) and (2) may seem odd at first sight, we are all familiar with them. An analogy might bring the point home. Consider your capacity to read aloud the sentence 'The Aguasabon Falls is a must-see attraction in Terrace Bay.' You would not be able to read aloud that sentence unless you believed that **that word** is pronounced [ði:] – where '**that word**' refers to the first word of the sentence above. So, when you read the sentence aloud, you surely believe at that moment that **that word** is pronounced [ði:]. Of course, you do not *think* or *judge* that **that word** is pronounced [ði:] when you read the sentence aloud – you just *believe* it. Moreover, of course, you do not have any awareness of your belief that **that word** is pronounced [ði:] when you read the sentence aloud,



nor do you remember having had that belief in the aftermath. The belief, as it were, entirely stays in the background (and disappears once the word ‘the’ gets out of sight). Nonetheless, it is (or was) there. The same applies to the beliefs that correspond to premises (1) and (2): they are there as long as the first-order judgment is performed, but entirely stay in the background (and vanish once the mental utterance ‘Toronto is located in Ontario’ dies away).

## 5. Résumé

The main question of the paper was: ‘Why is the transition from first- to second-order judgment rational from the perspective of the subject who performs that transition?’ Byrne did not provide an answer; he only told us that the transition is highly reliable. Boyle correctly identified this weakness in Byrne’s account, but his counterproposal had some shortcomings. According to Boyle, the transition is rational from the subject’s point of view because it involves a shift of attention to some mental fact of which the subject was already tacitly aware. However, the notion of a shift of attention from the external to the mental realm bears the danger of losing the idea of transparency. Therefore, I outlined an alternative. According to this alternative, the transition from first- to second-order judgment is rationally permissible in virtue of a justifying argument that starts out from a premise referring to the mental utterance emitted during the act of judging. Thus, Boyle is right when he claims, *pace* Byrne, that we do not infer that we believe that  $p$  from the premise that  $p$ . However, Boyle is wrong when he thinks that we are permanently tacitly aware of our beliefs. Rather, whenever we make a judgment to the effect that  $p$ , there is a set of background beliefs by virtue of which we are inferentially justified in judging that we believe that  $p$ . In short: we have no tacit permanent self-awareness of our first-order beliefs, but we do have permanent inferential justification for forming second-order judgments.<sup>18</sup>

## Notes

1. I deal with self-knowledge of non-doxastic attitudes such as wishes, desires, and intentions in Barz (2015). For more on self-knowledge of one’s visual experiences, see Barz (2014).
2. Cf. Gertler (2011), Cassam (2014), Finkelstein (2012).
3. I owe this label to Andi Müller.
4. See Peacocke (1998, 90) for an example that is often assumed to cast in an unfavorable light the thesis that judgment is sufficient for belief.
5. Cf. Byrne (2005, 95, 2018, 74–98). See also Dretske (2003, 2), Evans (1982, 231), Gallois (1996, 47), Martin (1998, 110) and Moran (2003, 413).
6. My formulation slightly deviates from Byrne’s as he uses ‘believe’ instead of ‘judge.’ However, in his (2005), footnote 22, Byrne himself admits that using ‘judge’ would be better.
7. Cf. BonJour (1980, 63).

8. Maybe this is not quite fair to Boyle's view – for Boyle might claim that knowledge is not a form of awareness but a kind of ability. See Marcus (2016) and Campbell (2018) for defenses of Boyle's reflectivism along these lines. However, since my chief objection to Boyle's account does not depend on whether knowledge is a form of awareness or not, I do not pursue the matter any further here. Thanks to David Hunter who drew my attention to this point.
9. Note, again, that the assumption that there are mental words is neutral concerning the question of how mental content is physically encoded. Thus, thinking a mental sentence is *not* the same as 'tokening a string of symbols of the language of thought' in Fodor's sense. Thanks to Peter Kuhn for encouraging me to be clearer on this point.
10. I owe this objection to Henning Lütje.
11. To forestall possible misunderstandings, let me emphasize that the justification in question licenses the subject merely to believe that the author of the utterance – whoever that may be – is sincere. It does *not* license the subject to believe that the author of the utterance is identical to herself. Thus, the argument cannot directly proceed to the conclusion 'I believe that Toronto is in Ontario,' but needs premise (2) for this purpose.
12. One might object that it is conceptually impossible to sincerely assert *p* without knowing that it is oneself who asserts *p*. Thus, nobody can sincerely assert that *p* but, for example, wonder whether it is him or her who asserts *p*. However, it seems that my description of the case presupposes that it is possible to sincerely assert that Toronto is located in Ontario without knowing that it is oneself who asserts it. Hence, there is a problem. In my opinion, this objection is based on a misunderstanding. Note that I do not describe a point in time of a real subject's mental life here. Instead, I describe the logical stage of the justifying argument I am about to formulate. Compare: from a psychological point of view, it might be impossible to believe 'A is a bachelor' without believing that A is male – for believing the first proposition without the second would show that one does not master the concept 'bachelor.' However, from a logical point of view, the premise 'A is a bachelor' does not license one to conclude that A is male; the further premise 'All bachelors are male' is needed. So, the fact that one is not licensed to conclude (without further ado) from 'A is a bachelor' that A is male does not imply that it is possible to believe that A is a bachelor without believing that A is male. Similarly, in the case at hand, the fact that premise (1) does not license the subject to conclude that she is the one who asserts that Toronto is located in Ontario does not imply that it is possible to sincerely assert that Toronto is located in Ontario without knowing that it is oneself who asserts it. Thanks to Eric Marcus for prompting me to reconsider and improve my thoughts on this matter.
13. I owe this and the following objection to Sarah Paul. I thank her for pressing me on these points.
14. Cf. Enoch and Schechter (2008, 568).
15. This reply might seem perplexing at first sight. One may say: 'Given that the Enoch and Schechter strategy is a form of externalism, why do you adopt it when justifying the first conjunct of premise (2)? Isn't that detrimental to the purpose of explaining why transitions from first- to second-order judgments are rational from the subject's point of view?' Answer: no, it isn't. Assume that the first conjunct of premise (2) is justified in the Enoch and Schechter way and that

the subject does not believe that believing the first conjunct of premise (2) is indispensable for being rational. This does not imply that the subject does not believe the first conjunct of premise (2). On the contrary, it is perfectly possible that the subject does believe the first conjunct of premise (2), though she has no grasp of its justifier. Note that, to be justified in believing *B* on the basis of *E*, one needs to be justified in believing *E*, but one does not need to have cognitive access to *E*'s justifier. Hence, the fact that the justifier of the first conjunct of premise (2) might be inaccessible from the subject's point of view does not mean that the justification of the conclusion 'I believe that Toronto is located in Ontario' is likewise inaccessible. In short: even if the subject does not know why she is justified in believing the premise, she is in a position to know why she is entitled to draw the conclusion. Thus, use of the Enoch and Schechter strategy in connection with the first conjunct of premise (2) is not detrimental to the purpose of explaining why transitions from first- to second-order judgments are rational from the subject's point of view.

16. Byrne (2018, 14). See also Byrne (2005, 92).
17. I consider this to be the received opinion about inferential justification. If the subject were required to actually entertain and accept all propositions that constitute the evidence justifying her belief, then virtually no belief would ever be inferentially justified. Cf. Fumerton (1976, 566).
18. An earlier version of this paper was presented at the workshop 'Transparency and Apperception' organized by Boris Hennig, David Hunter, and Thomas Land at Ryerson University, Toronto, in Mai 2018. Thanks to the audience on that occasion for helpful discussion, especially David Barnett, Boris Hennig, Ulf Hlobil, David Hunter, Thomas Khurana, Thomas Land, Eric Marcus, Sarah Paul, Gurpreet Rattan, Houston Smit, and Jonathan Way. For extremely helpful comments on a previous draft many thanks to Philipp Hey, David Hunter, Peter Kuhn, Andi Müller, Henning Lütje, and Sarah Paul. Finally, special thanks to Mark Davies (who knows why).

## References

- Barz, W. 2014. "Introspection as a Game of Make-Believe." *Theoria* 80: 350–367. doi:10.1111/theo.2014.80.issue-4.
- Barz, W. 2015. "Transparent Introspection of Wishes." *Philosophical Studies* 172: 1993–2023. doi:10.1007/s11098-014-0386-9.
- Boghossian, P. 1989. "Content and Self-Knowledge." *Philosophical Topics* 17: 5–26. doi:10.5840/philtopics198917110.
- BonJour, L. 1980. "Externalist Theories of Empirical Knowledge." *Midwest Studies in Philosophy* 5: 53–73. doi:10.1111/j.1475-4975.1980.tb00396.x.
- Boyle, M. 2011. "Transparent Self-Knowledge." *Proceedings of the Aristotelian Society Supplementary Volume* 85: 223–241. doi:10.1111/j.1467-8349.2011.00204.x.
- Byrne, A. 2005. "Introspection." *Philosophical Topics* 33: 79–104. doi:10.5840/philtopics20053312.
- Byrne, A. 2018. *Transparency and Self-Knowledge*. Oxford: Oxford University Press.
- Campbell, L. 2018. "Self-Knowledge, Belief, Ability (And Agency?)." *Philosophical Explorations* 21: 333–349. doi:10.1080/13869795.2018.1426779.
- Cassam, Q. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press.

- Dretske, F. 2003. "How Do You Know You are Not a Zombie?" In *Privileged Access*, edited by B. Gertler, 1–13. Aldershot: Ashgate.
- Enoch, D., and J. Schechter. 2008. "How Are Basic Belief-Forming Methods Justified?" *Philosophy and Phenomenological Research* 76: 547–579. doi:10.1111/phpr.2008.76.issue-3.
- Evans, G. 1982. *The Varieties of Reference*. Oxford: Clarendon Press.
- Finkelstein, D. 2012. "From Transparency to Expressivism." In *Rethinking Epistemology – Volume 2*, edited by G. Abel and J. Conant, 101–118. Berlin, New York: de Gruyter.
- Fumerton, R. A. 1976. "Inferential Justification and Empiricism." *The Journal of Philosophy* 73: 557–569. doi:10.2307/2025616.
- Gallois, A. 1996. *The World Without, the Mind Within*. Cambridge: Cambridge University Press.
- Gertler, B. 2011. "Self-Knowledge and the Transparency of Belief." In *Self-Knowledge*, edited by A. Hatzimoysis, 125–145. Oxford: Oxford University Press.
- Marcus, E. 2016. "To Believe Is to Know that You Believe." *Dialectica* 70: 375–405. doi:10.1111/dltc.v70.3.
- Martin, M. 1998. "An Eye Directed Outward." In *Knowing Our Own Minds*, edited by C. Wright, B. Smith, and C. Macdonald, 99–121. Oxford: Oxford University Press.
- Moran, R. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.
- Moran, R. 2003. "Responses to O'Brien and Shoemaker." *European Journal of Philosophy* 11: 402–419. doi: 10.1111/1468-0378.00193.
- Peacocke, C. 1998. "Conscious Attitudes, Attention, and Self-Knowledge." In *Knowing Our Own Minds*, edited by C. Wright, B. Smith, and C. Macdonald, 63–98. Oxford: Oxford University Press.
- Ryle, G. 1949. *The Concept of Mind*. London: Hutchinson's.
- Shah, N., and D. Velleman. 2005. "Doxastic Deliberation." *The Philosophical Review* 114: 497–534. doi:10.1215/00318108-114-4-497.
- Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
- Williamson, T. 2000. *Knowledge and Its Limits*. Oxford: Clarendon Press.