# Is self-knowledge of belief transparent?

**Abstract:** This article presents a dilemma for the view that conscious awareness of our own beliefs can be achieved transparently, entirely though consideration of the content or subject-matter of those beliefs. Either the judgments about that subject matter are expressed in inner or outer speech, in which case transparency is violated. Or they figure among the contents of working memory, in which case the account makes predictions that are belied by an extensive literature in social psychology.

Following some seminal remarks by Evans (1982), it has been argued that our knowledge of our own beliefs is *transparent* (Byrne 2005, 2018; Fernández 2013). To have transparent knowledge of one's own beliefs is to know of them in the same outward-directed way that one learns of the subject-matter of those beliefs. To learn what one believes, on this account, one looks outward to the world, not inward to our own minds. If one has transparent knowledge that one believes that p, then the basis for that knowledge is facts bearing on the truth or falsity of *p*, not facts about one's own mind, nor facts about one's own circumstances or behavior. In what follows I will focus on the account provided by Byrne (2018), leaving open whether the arguments I offer will generalize to other versions of the transparency approach.

Byrne argues that it is a constraint on theories of self-knowledge in general — and on a transparency account of self-knowledge of belief in particular — that they should be able to explain how self-knowledge is both *privileged* (more reliable than our knowledge of other people's beliefs) and *peculiar* (employing a method that can only be applied in the first person). In order to meet these constraints, what he suggests is that people employ the following self-attribution rule:

BEL: if *p*, believe that you believe that p.

However, since the rule might not be explicitly represented, but remain tacit in the inferential transitions that people make, an alternative formulation might be this:

BEL*: *p* → I believe that p.

It is obvious that the resulting method is peculiar. For the rule BEL is only reliable when employed in the first person. One cannot reason: *p*, so John believes that p. Indeed, the rule BEL* *can* only be employed in the first person. It is also obvious that knowledge acquired in this way is privileged. For the rules are (almost) self-verifying. In judging *p*, one *is* believing that p. So the conclusion that I believe that p is guaranteed to be true, and will amount to knowledge in all but the few rare cases where something

would have caused me to judge that I believe that p anyway, whether or not I was currently judging that p. So the real question is whether BEL and BEL* are transparent.[1]

Transparent self-knowledge was initially claimed with respect to one's awareness of one's own experiential states (Moore 1903; Evans 1982; Dretske 1994), and it is here that it is most readily defensible, I suggest. Everything required for knowledge that one is *visually experiencing* is present in the represented contents of one's experience. This is because visual perception (and visual imagery) always have contents that are unique to vision. Any visual experience will include a visual representation of color, for example (included in which is representation of black, white, and shades of grey). But no other sensory system can detect color. We can thus achieve self-knowledge of sight with the following transparency-securing rule (in which the premise needn't be a propositional representation of color, but rather the nonconceptual color-content of what is, in fact, a visual experience of some sort):

SEE: color of something there → I am seeing / visually experiencing something there.[2]

A similar rule can undergird one's knowledge that one is hearing. Since timbre, pitch, and loudness are properties that are uniquely represented in audition, one can employ a rule that moves from the represented contents of hearing to the knowledge that one is hearing. Likewise, one's knowledge that one is touching, tasting, or smelling something can plausibly be entirely world-directed.

One reason why the self-attribution rules for sight and hearing are straightforwardly transparent, whereas the rules BEL and BEL* aren't, is that the former operate entirely on the *content* of the mental states in question, without needing to be sensitive to one's *attitude toward* that content or to one's *mode* of content possession. Seeing a tree fall in the forest is a different mode of apprehending the tree falling than is hearing a tree fall. While aspects of the content are shared (*a tree is falling*), other aspects are unique to the mode in question (sight versus hearing), and the latter are sufficient for the

---

[1] Boyle (2011) argues that a rule like BEL cannot issue in knowledge because it fails to meet internalist rationality constraints on knowledge. I propose to set this objection aside. My concern is with transparency, not with the nature of knowledge.

[2] Whether one can have transparent knowledge of the *distinction* between visual perception and visual imagery is one of the questions Byrne (2018) addresses. I won't pursue that here, since my present goal is only to draw a contrast between the transparency of visual experience and the (alleged) transparency of belief. For the same reason, I here glide over all the qualifications and complexity inherent in Byrne's actual treatment of SEE. Notice, too, that if all seeing *is* a form of believing, as Byrne argues, then one can straightforwardly move from the content of visual perception to the conclusion that one is believing that content. Our focus here, however (as in Byrne's own discussion of BEL), is on beliefs that aren't directly grounded in current perception.

self-attribution rules to apply. The rule for BEL, in contrast, can't just take any propositional content as its antecedent. This is because one can imagine that p, suppose that p, doubt whether p, and so on; and in none of these cases can one transition from the proposition *p* to the conclusion that one believes that p. Rather, as Byrne himself makes clear, the rule BEL operates on current judgments as input. It is only when one is *judging* that p that one can deploy the rule, and hence judge that one believes that p.

Of course, it is no part of Byrne's view that one knows to apply BEL by being aware that one is judging that p. On the contrary, appealing to the use of BEL is supposed to explain how one comes to be aware of one's judgments in the first place. But the challenge for Byrne is to explain how a rule like BEL can be uniquely (or at least reliably) sensitive to the presence of judgments in contrast with other propositional attitudes, while at the same time preserving transparency by being entirely world-directed, focused only on the subject-matter of the belief. How does the fact of judging that p trigger an application of BEL independently of one's awareness that one is judging? For the content of the judgment alone is insufficient.[3]

Plainly it can't be acceptable for Byrne to say that the applicability of BEL depends on prior representations of one's mental states, nor on previous context or behavior. For example, if the way that one determines that a token of the content *p* can warrant an application of BEL is that it comes immediately to mind in response to someone asking one a direct question ("Do you think that *p*?"), then the application of BEL, here, won't be transparent in the relevant sense. For it will tacitly appeal to a principle of judgment-attribution, such as, "If the thought that p comes swiftly to mind when one is asked whether p, then one is judging that p". The result is that applying BEL wouldn't be entirely transparent; rather, it would depend on an inference from a generalization about how our minds work.

One solution to the challenge of explaining how BEL can be *sensitive* to judgments without depending on prior awareness of those judgments would be to suggest that the attitude of judging is carried by a formal property of the propositional representation involved in the judgment. Compare assertion. Whenever one utters a sentence that has indicative form, it is at the same time legitimate to preface one's utterance of the sentence with "I think that…". If one utters "It is raining" unadorned, without any preceding phrase like "suppose…" or "imagine…" and without any subsequent qualifier like

[3] There might, of course, be local ("sub-personal") regions of the mind where a rule like BEL is applied, issuing metacognitive representations of belief that are used for local processing purposes. But this isn't Byrne's target, nor is it the target of discussions of self-knowledge generally. Our goal is to explain personal-level knowledge of our own beliefs that can enter into reflective reasoning, decision making, and verbal report.

"in my imagination", then one can say instead, "I think it is raining." Indeed, English and many other languages employ a rule somewhat like BEL, permitting one to preface any assertoric utterance with "I think". The result, however, is not generally reckoned to attribute a belief to oneself as part of the speaker-meaning the utterance (as opposed to its literal semantics). If you ask me about the weather and I reply, "I think it is raining", the topic remains the weather, not my mental states. This sort of indirect-assertion usage of "think" is quite common in ordinary discourse (Shatz et al. 1983; Diessel & Tomasello 2001; Simons 2007).

It is plain, moreover, that indirect-assertion uses of "think" can't do the work of BEL. In part this is because the indicative form is an unreliable indicator that the content of the sentence is being asserted. Any indicative sentence can be used ironically, for example, to communicate the opposite of what it literally means. One can say, "It will be sunny today" and mean — and be taken by one's hearers to mean — that the weather will be terrible. Yet one can equally utter ironically, "I think it will be sunny today", also meaning that the weather will be terrible. In addition, the indicative form is an unreliable indicator that the speaker is expressing a judgment of any sort. One can say, "It will be sunny today" as a joke, or to ask a question; and the same is true when the sentence is prefaced with "I think".

Perhaps, however, there is a judgment-marker in the language of thought, as Rey (2013) suggests. Perhaps what fixes the judgmental role of any given propositional representation (fixing long-term memory, playing the guidance role in practical reasoning, and guiding linguistic assertion) is a formal marker of some sort attaching to the representation. The judgment-marker could either be explicit in the representational vehicle of the judgment, or it might perhaps be implicit in the neural network from which the content *p* emanates. The BEL rule could then be constructed to respond the presence of such a marker, and only such a marker. Nevertheless, the only *content* entertained in the antecedent of BEL would be world-directed, and hence transparency in self-knowledge of belief would be insured.

The sort of self-knowledge that forms our target is personal (as opposed to sub-personal) in nature, of course. What people are seeking to explain and understand is how one can have knowledge of beliefs that are active at the personal level, informing inferences and decisions, and being expressible in speech. In short, the judgments that BEL is supposed to give us knowledge of are those that are "access conscious" in just this sense: they are the ones that can enter into reflective reasoning and decision making, and that can issue in verbal statements. It is widely assumed, however, both in psychology (Dehaene 2014) and philosophy (Block 2007, 2011), that access consciousness can generally be equated with the contents of working memory. If we assume that this is so, then much will then turn

on the nature of working memory and its contents.

One possible view is that amodal (non-sensory) propositional attitudes themselves can figure in working memory. And in that case a view much like that sketched above would be possible. One's judgment that p when tokened in working memory with a sub-personal marker of some sort that identifies it as a judgment could trigger the application (when relevant) of the inference rule BEL* to issue in the judgment that one believes that p. The resulting belief would be transparently acquired, since one's entire focus throughout could be on the subject matter of p, without any appeal to one's own behavior or circumstances. I shall suggest shortly, however, that there is evidence against any such marker-dependent operation of the rules BEL or BEL*.

There is an alternative view of the contents of working memory, however, which has been claimed by Carruthers (2015) to be supported by the available scientific evidence. This is that working-memory contents are always sensory-based, and cannot include tokens of amodal attitudes like judgments, decisions, or intentions.[4] Those contents can look judgment *like*, however, when they contain items of assertoric inner speech. Asked for the name of the capital of France, the word "Paris" might emerge in working memory, and is then naturally taken as expressing the judgment that Paris is the capital of France. But a rehearsed utterance of "Paris is the capital of France" is not itself a judgment, of course (any more than is the utterance of those words out loud), but at best *expresses* a judgment. And then it seems that the result is that transparency is undermined.

For instance, suppose that the attribution-rule employed is one that takes assertoric utterances or rehearsed inner-speech assertions as input, thus:

BEL*/ASSERT: "[assertion] *p*" → I believe that p.

There are then two problems with such an account. One is that it presupposes knowledge of the intention with which the utterance is made. How is one to know — transparently — that in uttering or rehearsing an utterance of "p" one intends to assert that p, rather than the opposite, or instead of merely musing about p, or whatever? Indeed, Byrne (2018) offers no account of one's knowledge of intentions-in-acting (only of intentions for the future, knowledge of which he thinks can be derived from a certain sort of belief about the future). But secondly, and more obviously, transparency is violated because the input to the rule isn't confined to facts about the subject matter of *p*, but rather facts about

[4] Interestingly, Byrne, too, seems committed to such a view. In the final chapter of his book (2018, pp.207-8) he discusses and attempts to explain away evidence of unsymbolized imageless conscious thinking, concluding that if real, this could be the downfall of his transparency account of self-knowledge.

one's own behavior, namely that one has just asserted that p. If one has to know what one is doing in order to know what one believes, then one's access to one's own beliefs isn't transparent.

It seems, then, that Byrne (2018) must commit to some substantial claims about the nature of working memory and the contents of access consciousness. He must claim that they include amodal judgments that are somehow marked as such, enabling them (and in general, only them) to be taken as input by the rules BEL or BEL*. However, such an account makes a clear prediction. If the contents of working memory include judgments that are transparently marked as such, then one should immediately be aware of one's error if one attributes to oneself a *different* judgment. This prediction is known to be false, however, on well-established experimental grounds.

Decades of work in social psychology have demonstrated that people's overt expressions of their judgments can be pushed around by contextual factors, in such a way that what they say is something other than they actually believe (Cialdini & Petty 1981; Kunda 1999; Moskowitz 2005). But this happens unwittingly, outside of people's awareness. Consider the counter-attitudinal essay paradigm, for instance (Elliot & Devine 1994; Simon et al. 1995; Gosling et al. 2006). If people are manipulated into writing an essay supporting something that is the opposite of what they believe, while at the same time they are manipulated into feeling that they are choosing to do so *freely*, they will thereafter shift their expression of their attitude in the direction of the opposing view they have just defended, seemingly in order to make themselves feel better about what they have done. This finding is robust, and the effects are often large ones. But they have nothing to do with the quality of the arguments that people offer — participants haven't actually convinced themselves of the truth of the view they have defended. For if given the opportunity to make themselves feel better in some other way first — for example, by denying the importance of the issue, or by denying responsibility — their expression of their previous view is unchanged.

If one assumes that the question, "p?" will evoke one's belief that p when one has such a belief, issuing in the judgment that p, then in the counter-attitudinal essay paradigm the concluding question will evoke a judgment that conflicts with what one has just written. And if that judgment figures in working memory and is transparently marked as such, then one should immediately be aware that one is lying if one asserts the opposite. But plainly people have no such awareness. If they did, then one would expect them to feel worse, not better.

Admittedly, not all expressions of judgment can be manipulated in this sort of way. The findings only concern beliefs that *matter* to one (where one can be induced to feel bad by arguing the opposite) or that matter to others (where desires for social acceptance or social approval can induce one

unwittingly to say something other than one actually believes). Beliefs on mundane matters are likely to be immune from such influences. If one is asked what one believes to be the capital of France, the judgment that it is Paris is likely to be straightforwardly expressed. But this is of no help to Byrne or other transparency-theorists. For the rule BEL isn't supposed to be restricted to mundane beliefs only. And even if it were, that in itself would be enough to destroy transparency. For in applying the rule one would then need to appeal to more than just facts that are relevant to the truth of *p*. One would also need to know whether *p* is important to oneself or to other people.

Byrne (2018) is caught in a dilemma, then. On the one hand he can allow, with Carruthers (2015) and as he himself seems to accept in another context, that the access-conscious contents of working memory are always sensory-based ones, being confined to such things as visual images and inner speech. But in that case one's knowledge of one's own judgments cannot be transparent. It will rely, rather, on knowledge of things other than the subject-matter of those judgments, such as one's own rehearsed assertions or visual images. Or Byrne can claim, on the other hand, that judgments themselves can figure in working memory, in some way transparently marked as such. But in that case he runs afoul of the extensive social-psychology literature showing that we can easily be manipulated into asserting things other than we think, but without realizing it, and yet in circumstances where our real judgments would almost certainly have been activated.

**References**

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30, 481-499.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 12, 567-575.

Boyle, M. (2011). Transparent self-knowledge, *Aristotelian Society Supplementary Volume*, 85, 223-241.

Byrne, A. (2005). Introspection. *Philosophical Topics*, 33, 79-104.

Byrne, A. (2018). *Transparency and Self-Knowledge*. Oxford University Press.

Carruthers, P. (2015). *The Centered Mind: What the science of working memory shows us about the nature of human thought*. Oxford University Press.

Cialdini, R. & Petty, R. (1981). Anticipatory opinion effects. In R. Petty, T. Ostrom, & T. Brock (eds.), *Cognitive Responses to Persuasion*, Erlbaum.

Dehaene, S. (2014). *Consciousness and the Brain*. Viking Press.

Diessel, H. & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-

based analysis. *Cognitive Linguistics*, 12, 97-141.

Dretske, F. (1994). Introspection. *Proceedings of the Aristotelian Society*, 94, 263-278.

Elliot, A. & Devine, P. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67, 382-394.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.

Gosling, P., Denizeau, M., & Oberlé, D. (2006). Denial of responsibility: A new mode of dissonance reduction. *Journal of Personality and Social Psychology*, 90, 722-733.

Fernández, J. (2013). *Transparent Minds: A study of self-knowledge*. Oxford University Press.

Moore, G.E. (1903). The refutation of idealism. *Mind*, 7, 1-30.

Moskowitz, G. (2005). *Social Cognition*. Guilford Press.

Kunda, Z. (1999). *Social Cognition*. MIT Press.

Rey, G. (2013). We are not all "self-blind": A defense of a modest introspectionism. *Mind & Language*, 28, 259-285.

Shatz, M., Wellman, H., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, 14, 301-321.

Simon, L., Greenberg, J., & Brehm, J. (1995). Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology*, 68, 247-260.

Simons, M. (2007). Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117, 1034-1056.