Received: 10 July 2017 | Accepted: 21 October 2017

DOI: 10.1111/ejop.12318

WILEY European Journal of Philosophy

ORIGINAL ARTICLE

Introspection, mindreading, and the transparency of belief

Uwe Peters[1,2]

[1] Centre for Logic and Analytic Philosophy, Institute of Philosophy, KU Leuven, Kardinaal Mercierplein 2, 3000, Leuven, Belgium

[2] Department of Economics, University College London, Gower St, Kings Cross, London WC1E 6BT, UK

Correspondence
Uwe Peters, Centre for Logic and Analytic Philosophy, Institute of Philosophy, KU Leuven, Kardinaal Mercierplein 2, 3000 Leuven, Belgium.
Email: uwe.peters@kuleuven.be

Abstract

This paper explores the nature of self-knowledge of beliefs by investigating the relationship between self-knowledge of beliefs and one's knowledge of other people's beliefs. It introduces and defends a new account of self-knowledge of beliefs according to which this type of knowledge is developmentally interconnected with and dependent on resources already used for acquiring knowledge of other people's beliefs, which is inferential in nature. But when these resources are applied to oneself, one attains and subsequently frequently uses a method for acquiring knowledge of beliefs that is non-inferential in nature. The paper argues that this account is preferable to some of the most common empirically motivated theories of self-knowledge of beliefs and explains the origin of the widely discussed phenomenon that our own beliefs are often *transparent* to us in that we can determine whether we believe that *p* simply by settling whether *p* is the case.

1 | INTRODUCTION

What is the nature of and developmental relationship between our knowledge of our own beliefs (henceforth *self-knowledge of beliefs*) and our knowledge of other people's beliefs (henceforth *other-knowledge of beliefs*)? Two different proposals are often distinguished. I shall call them the *asymmetry view* and the *symmetry view*.

According to the asymmetry view, self-knowledge of beliefs is independent from other-knowledge of them. It is developmentally prior to and typically involves a different faculty than other-knowledge of beliefs. Unlike other-knowledge of beliefs, self-knowledge of them is normally[1] acquired without observation, interpretation, or inferences (Armstrong, 1999; Goldman, 2006; Harris, 1989; Nichols & Stich, 2003).

In contrast, the symmetry view holds that self-knowledge of beliefs is dependent on other-knowledge of them. It is developmentally posterior to or simultaneous with the latter and both involve the operation of the same cognitive systems. Self-knowledge of beliefs is, just as other-knowledge of them, dependent on observation or interpretation, more generally, on inferences (Carruthers, 2011; Cassam, 2017; Dennett, 1992; Dretske, 2003; Gopnik, 1993; Ryle, 1949; Stephens & Graham, 2000).[2]

In the following, I propose a novel account of self-knowledge of beliefs that combines elements of both the asymmetry view and the symmetry view. I argue that self-knowledge of beliefs is developmentally posterior to or

simultaneous with other-knowledge of them and deploys resources already used for acquiring other-knowledge of beliefs, which is inferential in nature. But when these resources are applied to oneself, one attains and subsequently frequently uses a method for acquiring self-knowledge of beliefs that is non-inferential in nature.

I contend that this account is preferable to some of the most common empirically motivated versions of both the asymmetry view and the symmetry view. It also captures and explains the origin of the widely discussed phenomenon that our own beliefs are often *transparent* to us in that we can come to know whether we believe that *p* simply by determining whether *p* is the case[3] (Byrne, 2011; Evans, 1982; Moran, 2001).

In Sections 2 and 3, I set the scene for the main discussion. I begin, in Section 2, by introducing two influential empirically supported versions of the asymmetry view, namely Shaun Nichols and Stephen Stich's (2003) theory, and Alvin Goldman's (2006) account. In Section 3, I turn to the symmetry view and introduce Peter Carruthers' (2011) recent version of it. In critical response to these proposals, I then develop my own account in Section 4. In Section 5, I mention and rebut two objections to the account, before, in Section 6, arguing that the belief self-ascriptions that it covers qualify as non-inferential self-knowledge. Section 7 summarizes and concludes the discussion.

2 | THE ASYMMETRY VIEW

One prominent empirically supported version of the asymmetry view is Nichols and Stich's (2003) account. Nichols and Stich argue that belief self-ascriptions, that is, second-order beliefs with which one ascribes a belief to oneself, are typically the product of a simple detection process, a

> *Monitoring Mechanism (MM) (or perhaps a set of mechanisms) that, when activated, takes the representation* p *in the Belief Box as input and produces the representation* I believe that p *as output. […] To produce representations of one's own beliefs, the Monitoring Mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form:* I believe that ___, *and then place the new representations back in the Belief Box. (Nichols & Stich, 2003: 161)*[4]

The MM is thought to be a sub-personal mechanism that is distinct from the "mindreading" or "theory-of-mind" system, which is the cognitive capacity that enables us to determine other people's mental states including their propositional attitudes (henceforth *PAs*) such as beliefs, desires, intentions, etc., via observation of, and inferences from people's behaviour or circumstances (ibid). Unlike the mindreading system, the MM produces PA self-ascriptions without any kind of observation or inference. Furthermore, while both the MM and the mindreading system can produce belief self-ascriptions, the MM is assumed to come online "significantly before" the mindreading system is "fully in place," and the latter only becomes active when one is trying to work out the causes of one's own behaviour, which involves reasoning about PAs (Nichols & Stich, 2003: 163). The MM and the mindreading system are viewed as independent, allowing for two-way dissociations in which the MM, and so the ability to detect one's own PAs, is intact, while the mindreading system, and so the ability to determine other people's PAs, is impaired and *vice versa*. Nichols and Stich (2003: 165f) argue that empirical studies from developmental psychology and data on autism and schizophrenia provide evidence for such dissociations and therewith for the MM theory. Studies indicate, Nichols and Stich hold, that both normally developing young children and autistic subjects have difficulty tracking other people's mental states but not their own. Reversely, in passivity-symptomatic schizophrenics, other-ascriptions of mental states appear intact but self-ascriptions of them are impaired (ibid).

I shall briefly return to some relevant data below. For now, I just want to note the following general issue with Nichols and Stich's account. The MM's task in forming belief self-ascriptions is to "copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that* ___, and then place the new representations back in the Belief Box" (Nichols & Stich, 2003: 161). Nichols and Stich maintain that the same holds for other PAs such as suppositions or doubts. The only aspect that then changes is the "Boxes" and the first part of the representation schema. The problem is that when the MM copies a particular mental representation, it still has to

determine from which "Box" it took the representation. Otherwise, it won't be able to produce correct PA ascriptions. So how can the MM tell whether the content of a mental representation *p* is *believed* as opposed to, say, supposed or doubted? Nichols and Stich do not say. Their view thus remains crucially incomplete.

One proposal to address this point is that mental representations might have attitude-type specific neural properties that the MM tracks and uses to determine whether a particular mental representation carries content that is believed rather than, say, supposed or doubted.

Goldman (2006) has developed a view of this kind. He agrees with Nichols and Stich that subjects have a "special method of accessing or detecting their current mental states," which is operative *prior* to and dissociable from the mechanism involved in determining other people's mental states (Goldman, 2006: 224). But unlike Nichols and Stich, Goldman maintains that this method involves "introspection," a quasi-perceptual, attention-dependent "inner recognition" of mental states based on their intrinsic properties (Goldman, 2006: 246). He thinks that just as perception, introspection relies on a transduction mechanism that is causally sensitive to a particular kind of input and produces, on the basis of that input, a particular kind of output. When it comes to determining mental-state types, the inputs are neural-activation patterns that are detected by a sub-personal attention network: a high level of activation in one class of neural cells "generates the introspective classification 'pain' […], a high level of activation in a different class of cells generates the introspective classification 'tickle'" and so on for other mental states including beliefs (Goldman, 2006: 252). Goldman takes his proposal to be capable of handling the problem of attitude-type identification that Nichols and Stich's account faces (ibid).

This is questionable, however. On Goldman's view, for a self-ascription of a belief that *p*, the belief that *p* must be tokened in first-order reasoning. Otherwise, the neural-activation pattern associated with the belief will not be present for the introspective attitude-type tracking mechanism to latch onto. Suppose, then, that my belief that *p* is now playing a role in my first-order reasoning and the neural activation at issue is occurring. How can the attitude-type detection mechanism take the neural-activation pattern to be indicative of my *believing* rather than, say, my supposing that *p*, and produce the self-ascription *I believe p* on its basis? The link between the neural-activation pattern and the second-order belief must somehow be established, for instance, via learning. The problem is that the connection will need to be established while I am only aware of the (for me) fact that *p* and while *p* figures in my first-order reasoning. Given that *p* might be the case even if I do not believe it, and I may believe that *p* even if *p* is not the case, how do I learn to link *p* and the self-ascription *I believe p*, and how does the introspective system set up the connection between the neural properties instantiated when *p* figures in my first-order thinking and the meta-representation *I believe p*? Goldman does not offer an explanation. His account too is thus in an important respect incomplete.

There is another problem with both Goldman's account and Nichols and Stich's view. Underlying both proposals is the thesis that self-ascriptions of beliefs are developmentally prior to other-ascriptions of them. Nichols and Stich (2003: 165f) and Goldman (2006: 236f) mention empirical findings from developmental psychology and data on autism and schizophrenia to support this thesis. But their interpretations of the data have been found to be flawed (Carruthers, 2011; Robbins, 2004), and there is in fact counterevidence to their priority thesis. For instance, if the thesis were right then a meta-analysis of developmental studies pertaining to children's understanding of beliefs should reveal temporal differences in the development of self-ascriptions of beliefs versus other-ascriptions of them. Yet, when Wellman, Cross, and Watson (2001) conducted an extensive meta-analysis of 178 studies on children's false-belief understanding that involved self- and other-conditions, they found that the

> *essential age trajectory for tasks requiring judgments of someone else's false belief is paralleled by an identical age trajectory for children's judgments of their own false beliefs. Young children, for example, are just as incorrect at attributing a false belief to themselves as they are at attributing it to others. (Wellman et al., 2001: 665)*

The apparent parallelism in the development of the ability to self-ascribe (false) beliefs and the ability to other-ascribe them challenges Nichols and Stich's, and Goldman's asymmetry view and supports the alternative proposal that there is a self-other symmetry in the development of belief ascriptions.[5]

3 | THE SYMMETRY VIEW

One well-defended empirically motivated version of the symmetry view is Carruthers' (2011) *interpretive sensory-access* (ISA) theory. It holds that self-ascriptions of PAs result from turning toward oneself the same typically unconsciously operating interpretive faculty that one already deploys to produce other-ascriptions of them, namely the mindreading system. The only difference is that in one's own case, the system has more sensory information available upon which to base its interpretation, for it can also access one's own mental imagery (e.g., inner speech), sensory-affective states, and the conceptual contents "bound into" them (Carruthers, 2011: 68f).

The ISA theory does not have the shortcomings that Nichols and Stich's and Goldman's views have. For instance, it avoids the problem of attitude-type identification that their views leave unresolved because it denies that, say, a belief that *p* is ascribed on the basis of a detection of a first-order representation about *p*. It proposes instead that the belief ascription is formed on the basis of an interpretation of an overt or mental-imagistic (e.g., inner speech) expression of the belief that *p*, and this interpretation rests on folk psychological generalizations that we already use to determine other people's beliefs (e.g., a subject's claims are expressions of her beliefs). In addition, unlike Nichols and Stich's and Goldman's views, the ISA theory is also well in line with a self-other symmetry in the development of belief ascriptions, for it holds that self- and other-ascriptions of PAs are formed by the same cognitive system.

To further motivate his proposal, Carruthers argues that in experimental settings, subjects often confabulate PAs for their own behaviour without awareness of doing so and while being under the impression that they have direct access to them (e.g., Brasil-Neto et al., 1992; Johansson, Hall, Sikstrom, & Olsson, 2005; Wegner & Wheatley, 1999). These findings, Carruthers continues, undermine our warrant for holding that we know our own PAs introspectively without interpreting ourselves because we may, just as the test subjects in these experiments, have the *impression* that we introspect them even though we in fact self-ascribe them interpretively (Carruthers, 2011: 325f).

At the same time, the data support the ISA theory, Carruthers holds. For in order to account for the findings, dual-method views, which hold that self-knowledge of PAs is at least sometimes non-interpretive, need to postulate two different faculties; one for non-interpretive and one for interpretive self-knowledge. The ISA theory assumes only one, an interpretive capacity. Since there is no positive support for the view that we have non-interpretive access to our own PAs apart from the intuition that we do, which is undermined by the data on unconscious confabulations, the ISA theory should be preferred, Carruthers (2011: 44f) argues, because it is simpler and explanatorily more general.

The ISA theory also has the advantage of providing an account of the evolution of self-ascriptions of PAs. For instance, we know that there is an other-directed, interpretation-dependent capacity to ascribe PAs and we have a good understanding of the pressure that could have led to its selection (e.g., the social-intelligence hypothesis, see Byrne & Whiten, 1997; Sterelny, 2012; Tomasello, 2014). However, we do not yet have a good grasp of the selection pressure that could have led to the evolution of an extra, non-interpretively operating mechanism for PA self-ascriptions, Carruthers (2011: 67f) maintains.[6] On the ISA view, the problem of explaining the evolution of such a mechanism disappears because there is no such mechanism to begin with. Our ability to self-ascribe PAs emerged from turning "inward"[7] an already existing outward-focused faculty (ibid).[8]

On the basis of empirical evidence (*inter alia* on PA confabulation, self-interpretation, the development of PA ascriptions, and working memory processing),[9] explanatory parsimony, and evolutionary considerations, Carruthers concludes that our access to our own (non-sensory/non-affective)[10] PAs such as, for instance, our own beliefs about the world is "*always interpretive* […], utilizing the same kinds of inferences (and many of the same sorts of data) that are employed when attributing attitudes to other people [emphasis added]" (Carruthers, 2011: 1).

There is a reason to be sceptical about this strong claim, however. For instance, the empirical evidence on attitude confabulation and self-interpretation that Carruthers uses to argue for the ISA theory does not show that we never have non-interpretive access to our own PAs. The data suggest at best that we often lack it. Carruthers' central argument for the generalization that we always only have interpretive access to our own PAs is that because we have no positive reason to believe that we ever have non-interpretive access to them, and because

the empirical findings suggest that we often lack it, the interpretation-only proposal should be preferred because it is more parsimonious than a dual-method alternative. In the remainder, I want to challenge this point by developing and defending an account of belief self-ascriptions that holds that when we apply some of the resources that we already employ to determine other people's beliefs to ourselves, then a non-interpretive and non-inferential method for forming belief self-ascriptions becomes available and is in fact frequently used. The account will combine elements of both the symmetry view and the asymmetry view.

4 | TOWARD A HYBRID VIEW

The account that I want to develop starts with the idea that during cognitive development, we apply to ourselves the following basic line of thought that we already utilize to determine other people's beliefs. In social interactions, we learn that one way in which we can determine whether a subject *S* believes that *p* is by simply asking her whether *p* is the case. Typically, S's response is an expression of her belief on whether *p* is the case.

Suppose that the mechanism for forming belief self-ascriptions recycles this idea. When we wonder whether we believe that *p* ourselves, then that idea will lead us to ask ourselves whether *p* is the case.

There is good reason to assume that we do in fact use that procedure. For many philosophers have observed, and take it to be a description of a common everyday life phenomenon, that the question "Do I believe *p*?" is often "'transparent' to, answered in the same way as, the outward-directed question as to the truth of *p* itself" (Moran, 2001: 61; see also Boyle, 2011; Byrne, 2005, 2011; Cassam, 2014, 2017; Evans, 1982; Peacocke, 1998). Call this phenomenon the *transparency of belief*.

Although there are different ways in which the transparency of belief has been construed, a common subsequent claim about it is that when upon determining whether *p* is the case, I conclude that *p*, I then transition directly, that is, without any kind of self-interpretation or intermediate inference, from *p* to the self-ascription *I believe p*[11] (Byrne, 2005, 2011; Gallois, 1996; Gordon, 2007; Moran, 2001). Call this the *transparency method* (TM) for forming belief self-ascriptions.

I am sympathetic to theories of belief self-ascriptions that appeal to TM, and I shall below develop a new TM-based account. Why might we need a new proposal of this kind?

The reason is that although it is a widely held view that we often use TM to form belief self-ascriptions, philosophers who subscribe to this view usually just take this as given and do not provide support for the claim that we link *p* with the self-ascription *I believe p* (see, e.g., Byrne, 2011: 207). The problem is that in the absence of an argument to that effect, advocates of other theories of belief self-ascriptions like Carruthers (2011) might deny that TM is used to determine whether one believes that *p*. He might hold that one does not directly transition from *p* to the belief self-ascription, at best, a transition from an inner-speech token "*p*" occurs that rests on an (unconscious) interpretive inference by the mindreading system.

In the following, I aim to block this response by introducing an account of belief self-ascriptions that explains the development of TM and supports the view that we do often use TM to determine our own beliefs rather than interpretive processing. I shall build up the account step by step. I begin by mentioning and motivating the two main assumptions that it rests on.

4.1 | Assumption 1

I shall take it that when one has the concept of belief, one understands the following principle linking action and belief

> *[ACTBEL] Subjects do not act directly on the basis of what is the case but on the basis of what they believe to be the case, which might but need not correspond to what is the case.*

Suppose one does *not* understand [ACTBEL]. Would one still count as having the concept of belief?

That this is arguably not the case is suggested by developmental data on children's performance in the verbal "false-belief task" (henceforth *FBT*), also known as the "Sally–Anne test" (Dennett, 1978; Wimmer & Perner, 1983). In the test, children are asked to observe two individuals, Sally and Anne, in a room with two boxes, B1 and B2. Sally has a marble and puts it into B1. She then leaves the room and, while she is outside, Anne takes the marble out of B1 and puts it into B2. Sally returns and the children are asked where she will look for the marble.

Studies found that 3-year-olds consistently say that she will go to B2, where the marble in fact is, which is an incorrect prediction. It is not until the age of 4 that neurotypical children are able to pass the test and say that Sally will look for the marble in B1, where she falsely believes it to be (Wellman et al., 2001). This is usually taken to show that it is not until the age of 4 that children acquire a full concept of belief (though a non-conceptual understanding of belief[12] is thought to be in place much earlier; see Low & Perner, 2012). In other words, in order to have this concept and form consistently correct belief ascriptions to other people across different FBT situations, children need to comprehend that in FBT contexts people do not act on the basis of what is the case but on the basis of what they believe to be the case.

Children could succeed in FBTs, if they assume that a subject *S* only acts on the basis of her beliefs in the *bad case*, i.e., in FBT scenarios, when a change in the world occurred without her noticing it, and in the *good case* (when the world didn't change without *S* noticing it) acts on the basis of facts. Notice, however, that to be able to consistently succeed in FBTs, children then still need to be able to tell whether *S* is in a good or bad case scenario, and to be able to tell either way, they will need to appeal to what *S* noticed and so takes to be the case. But that is just to say that they need to understand that even in the good case, subjects do not act *directly* on what is the case, i.e., without registering and taking it to be the case, but on their beliefs (facts can be reasons for acting too (Kolodny, 2005), but arguably only *indirectly* via the subject's beliefs and knowledge about them (Hornsby, 2008: 251f). Correspondingly, since the ability to consistently pass verbal FBTs is a prerequisite for having a full concept of belief, an understanding of [ACTBEL] too is part of possessing this concept.

The point that when one has the concept of belief one understands [ACTBEL] is the first assumption that the account of belief self-ascriptions which I want to develop rests on. I now turn to the second one.

4.2 | Assumption 2

I assume that when people think about the basis of their own action, including in FBT situations, they often find that they act directly on the basis of what is the case.

To support this, suppose you, a subject who has the concept of belief and so a grasp of [ACTBEL], are taking part in the FBT. Suppose that after correctly answering that Sally will search for the marble in B1, you are asked to retrieve the marble yourself. Since you know that it is in B2, you will go to B2. Suppose that when you are moving toward the box, you are interrupted and questioned as to why you are going to B2 rather than B1 to get the marble. It is fair to assume that you will respond by saying "because the marble is in B2, not B1." In other words, you will take it that you are acting directly on the basis of what is the case, namely on the basis that the marble is in B2.

Why do you take this to be your basis for acting even though you understand [ACTBEL], that is, that people do *not* act directly on the basis of what is the case? It might be proposed that the reason is that you construe that what is the case as that what you believe to be the case. If so, there would be no inconsistency with [ACTBEL]. But this proposal presupposes that you already have an insight into your beliefs. Since we are currently looking for an explanation of how you acquire this insight, we cannot presuppose that you already have it. In order to not presuppose what needs to be explained, let us henceforth suppose that in the envisaged FBT situation you do not yet classify that what is the case as that what you believe to be the case.

There is another answer to the question as to why you take the fact that the marble is in B2 to be your basis for acting even though you understand [ACTBEL]. Before introducing it, a bit of stage setting will be useful.

Note first that since you are the author of what you are doing and your acting is up to you, when you reflect on why you are acting in a particular way, you might settle the matter without appealing to evidence about yourself simply by choosing a reason and committing to acting on its basis. When you opt for this way of settling the issue, you do

not then *discover* the answer to the question as to why you are acting. Rather, you find out about it by selecting it (e.g., "I'm going to B2 because the marble is in that box.") and committing yourself to making it true.

Suppose this is how you do settle the issue. You will then re-enact the practical reasoning that initially (before you wondered about your acting) equipped you with a basis for acting and led you to go to B2. Since you know that the marble is in B2, the fact that it is in that box will for you be an excellent basis for acting. This is because, given that the marble is in B2, acting on the basis that it is will allow you to get it. Moreover, you are not forced to passively observe that this fact is a good basis for acting. Rather, you can turn it into your basis for acting by deciding to act on it. In fact, since you understand that acting on the basis that the marble is in B2 will help you get the marble, and you want to get the marble, you *do* decide to act on the basis that the marble is in B2 and commit to that course of action. This is confirmed by your verbal response to the above query about your acting.

Notice that when you decide to act on the basis that the marble is in B2, [ACTBEL] does not matter to you. It only becomes relevant to you when your goal is to work out *why* people are acting the way they do, not when you are trying to determine how to act yourself so as to get the marble. These are two different projects; the first is theoretical in nature, the second one practical.

Relatedly, for you to decide to act on the basis that the marble is in B2, you need not determine whether you believe or know that the marble is in B2. This would be entirely unnecessary. You just need to settle whether the marble is in the box. If it is then, from your own point of view, that fact will be a good basis for acting, because acting on the basis of what is the case will increase the probability that you get what you want. Indeed that the marble is in B2 will for you present a superior basis for acting if it is not just something someone believes. This is because you understand that beliefs might be false and so the likelihood of success in your interacting with the world is diminished if you act on the basis of what someone believes, rather than the facts. There is thus good ground to hold that even though you understand [ACTBEL], you will in the situation at hand, upon concluding that the marble is in B2, decide to act on the basis that it is.

Suppose you do. If so, this will lead you to take the fact that the marble is in B2 to be your basis for acting, because you view yourself in that situation as the "designer" of what you are doing and, as Velleman (2006) puts it, the

> *designer of something is the one whose conception of the thing determines how it is, rather than* vice versa, *and determines this by a mechanism reliable enough to justify his confidence in that conception as an accurate representation. To be the designer of something is just to be the one whose conception of it has epistemic authority by virtue of being its cause rather than its concomitant or effect. (Velleman, 2006: 262)*

If we apply this to the case at hand, then when you answer the question as to why you are going to B2 by selecting the fact that the marble is in B2 as a basis for your acting (and as an answer to the question), you are thereby (by your own lights) designing your action so that it is an action based on the fact that the marble is in B2 rather than on something else. Your *forward-looking* view that you are going to that box because the marble is in it constitutes an insight into what you are doing because it is determinative of your acting in a way that you take to be sufficiently reliable to support your certainty about the truth of the thought. As a result, you have an insight into your action that precedes the relevant evidence and is independent from interpretation or introspection.

Empirically oriented philosophers, including advocates of the ISA theory, will be suspicious of this proposal. This is because as already noted in Section 3, various studies have revealed that when asked about the motives for their choices and actions *post hoc*, people frequently fail to correctly identify their actual reasons and unknowingly confabulate explanations (Hall, Johansson, Tärning, Sikström, & Deutgen, 2010; Johansson et al., 2005; Nisbett & Wilson, 1977). The view just introduced might seem inconsistent with the data.

However, this is not the case. The confabulation findings do not in fact undermine the proposal here because they pertain to situations in which people reflect on the basis for their action *after* they acted. The data are compatible with the view that subjects have the mentioned forward-looking insight into the basis of their intentional actions prior to and during the execution of the action.

In any case, my argument in what follows only relies on the assumption that when subjects are in situations such as the FBT thinking about the basis of their own action, they will often find that they are acting directly on the basis of a particular non-mental fact (e.g., on the basis that the marble is in B2). If they come to this (possibly mistaken) view via an unconscious self-interpretation by the mindreading system or as a result of confabulation, this will not pose a problem for my argument. All that matters is that they do come to that view, no matter how and whether it is correct. And that people do in fact do so, even if they don't yet have a grasp of their own belief or knowledge states, is supported by the preceding considerations.

I shall now use the two assumptions just motivated to introduce a proposal on how subjects can come to directly transition from *p* to the self-ascription *I believe p*. The key idea will be that the two cognitive elements that the two assumptions pertain to – i.e., a subject's understanding of [ACTBEL], and her insight into the basis of her own action – lead the subject to an inner conflict that prompts her, during her cognitive development, to re-contextualize *p* as that what she believes to be the case.

4.3 | How TM can arise

Consider again the scenario in which you are taking part in the FBT, are asked to retrieve the marble yourself, and probed about the basis of your subsequent acting. Suppose that when you are reflecting on the basis of your action, you come to the view that you are acting on the basis that the marble is in B2. You thereby arrive at the view that you are acting directly on the basis of what is the case.

This leads to the following problem. By assumption, you also have a grasp of [ACTBEL], according to which, people do *not* act directly on the basis of what is the case but on the basis of what they believe to be the case. Suppose that in line with the recently popular view that self-knowledge of PAs rests on applying to oneself the mindreading system and other resources already used to make sense of other people's mind and behaviour (Carruthers, 2011; Cassam, 2017; Williams & Happé, 2010), you apply [ACTBEL], which is such a resource, to yourself. If you apply [ACTBEL] to yourself, your insight that you are acting directly on the basis of what is the case is likely to create a tension: from your point of view, the question will arise as to why you yourself would act on the basis of what is the case even though, as the principle tells you, people do not do so.

In response to this tension, you could revise the assumption that [ACTBEL] is a *general* principle. You could hold that you, as opposed to others, *do* act on the basis of what is the case.

However, this response is unlikely. One reason is that other people will reinforce in you the view that you also, just as everyone else, do not act directly on the basis of what is the case but on the basis of what you believe to be the case. This might happen, for instance, when they see you act on a false belief yourself. In such a situation, you cannot easily deny the truth of people's claim that you too are acting on the basis of your beliefs. For you will not find, say, the marble that you are searching for where you take it to be, and this will be difficult to comprehend for you unless you acknowledge that others, when they tell you in that situation that you are not acting on the basis of what is the case, are in fact right. In such situations, social interactions will incline you to refrain from revising [ACTBEL] so that it excludes you.

There is a different and more likely way for you to resolve the tension at issue. Since you understand that in the context of the FBT, to get the marble, subjects go to the box where they *believe* it to be, there is for you a link between what is the case, on the one hand, and the concept of belief, on the other. Hence, for you, if you do not act on the basis of the *fact* that the marble is in B2, the alternative that is set by the context is that you are acting on the basis of your *belief* that the marble is in B2. This is likely to nudge you, more specifically, the cognitive system in you that is tracking the matter (i.e., the mindreading system) to resolve the tension between [ACTBEL] and your insight that you are acting on the basis of what is the case by rejecting the view that what is the case is the case as a matter of fact, and by accepting that it is only what is the case *for you*, that is, what you believe to be the case. In fact, re-conceptualizing that what is the case as that what is the case from your perspective is your only way of resolving the mentioned tension that coheres with what you know about [ACTBEL], your own basis for acting, and others' testimony about

yourself (when you are acting on a false belief). These considerations lend plausibility to the assumption that you are likely to implement the re-conceptualization at issue.

Once it has occurred, a simple and elegant method for forming belief self-ascriptions becomes available. This is because you (or the cognitive system in you tracking the matter) can then take it that in settling whether *p* is the case, you are in fact only settling whether *p* is the case for you, that is, whether *p* is what you believe. Hence, in order to determine whether you believe *p*, you can use the following principle for belief self-ascriptions

[BSA] If *p* is the case, then I believe *p*.[13]

At first glance, it might seem that [BSA] implies that you now take yourself (absurdly) to be omniscient. If so, then we would have reason to doubt that you ever use the rule.

However, this problematic implication does not in fact follow. The reason is that when you are able to pass the FBT and have the concept of belief, you distinguish between reality and belief, between a perspective-independent *p* and a perspective-dependent *p*. With the re-conceptualization of *p* as *p* for you, this distinction becomes merely refined. You retain the concept of a perspective-independent *p* and now distinguish between *p* being the case from your perspective (henceforth *p**), and *p* being the case independently from anyone's perspective (henceforth *p***). That is, when you re-conceptualize *p* as *p**, you retain the idea that *p* might be perspective-independently the case (*p***) even if it is not the case from your perspective (*p**), and reversely, *p* might be the case from your perspective (*p**) even if it is not the case perspective-independently (*p***). Hence, it does not follow that once you re-conceptualize *p* as *p** that you then accept that if *p* is the case perspective-independently (i.e., *p***), then you believe it, and so it does not follow from accepting [BSA] that you are (or take yourself to be) omniscient. Rather, you (or, again, the cognitive system in you tracking the matter and implementing the re-conceptualization) take it that if *p* is the case, then *p* is only what is the case for you, which means that it might not be the case for others or perspective-independently, and reversely *p* might be the case for others or perspective-independently, even if it is not the case for you. Notice that no belief self-ascription is already presupposed here in the suggestion that you conceptualize *p* as *p**, and so as what you believe. For the preceding offers a step-by-step account of what exactly might lead you to this conceptualization in the first place, namely the need to resolve the mentioned tension between different background beliefs.

It is time to relate the account of belief self-ascriptions just introduced to the discussion of TM at the beginning of the section, and to the asymmetry and symmetry views. There are three points to note.

First, [BSA] captures the transition involved in TM. The preceding offers an explanation of how TM might arise during one's upbringing. It thereby casts new light on the phenomenon of the transparency of belief, because even though much has been written on the transparency of belief, a *developmental* account of it has not yet been proposed. The account of TM that I'm introducing in this paper has also many advantages compared to extant theories of belief-self-ascriptions that appeal to the transparency of belief. I show this elsewhere (see Peters 2017b).

Second, in line with the asymmetry view, the proposal introduces a *non-interpretive* method for forming belief self-ascriptions, for if one follows [BSA], one will transition directly from *p*, conceptualized as *p**, rather than from, say, a linguistic expression of one's belief that *p*, to the self-ascription *I believe p*. No interpretation of oneself or one's circumstances is involved.

Finally, the account coheres well with the empirical studies indicating a self-other symmetry in the development of belief ascriptions. This is because, in line with the symmetry view, the account proposes that belief self-ascriptions rely on applying to oneself the mindreading system and other resources already used for other-ascriptions of beliefs. The first one of these resources was the idea that one way of working out whether *S* believes that *p* is to ask her whether *p* is the case. The account introduced proposes that during our cognitive development, we apply this initially other-directed point to ourselves. The transparency of belief, that is, the phenomenon that the question "Do I believe *p*?" can be "answered in the same way as, the outward-directed question as to the truth of *p* itself" (Moran, 2001: 61), then results. The second initially other-directed resource for determining people's beliefs that is redeployed for belief self-ascriptions was one's understanding of [ACTBEL]. The mindreading system already needs the principle to be able to

correctly determine other people's actions and beliefs in FBT situations. Since belief self-ascriptions are, according to the account introduced, the result of a recycling of resources used for other-ascriptions of beliefs, it is to be expected that the ability to produce other-ascriptions of beliefs should be online before or at the same time as the ability to form self-ascriptions of them, and this is what the relevant studies suggest (e.g., Wellman et al., 2001). Hence, the proposal introduced offers a new contribution to research on the transparency of belief by combining aspects of the symmetry view with aspects of the asymmetry view (as subjects can non-interpretively self-ascribe beliefs).

But what reason is there to assume that the just-introduced developmental TM account of belief self-ascriptions does not only capture a *possible* way of forming belief self-ascriptions but a method that we are in fact likely to frequently employ? In the next section, I provide a reason and further defend the account. I shall do so indirectly by considering two critical responses to the latter.

5 | OBJECTIONS

5.1 | Objection 1

If we assume, as the developmental TM account does, that belief self-ascriptions emerge from applying to oneself resources initially directed at others, then the ISA theory offers arguably a more unified account, because it implies that the cognitive system involved in the formation of PA ascriptions will in both cases, in the case of self- and other-ascriptions of beliefs, determine a subject's beliefs via a process of interpretation. What basis is there for rejecting this view and for endorsing the developmental TM account?

The reason harks back to the transparency of belief, the phenomenon that when we are wondering whether we believe that *p*, we proceed to settle whether *p* is the case. It is widely accepted in philosophy that the phenomenon is real (e.g., Byrne, 2011; Cassam, 2014, 2017; Evans, 1982; Fernández, 2013; Moran, 2001). A theory of self-knowledge of PAs should thus have an account of it. Here is the ISA theory's. Upon wondering whether I believe that *p*, I ask myself whether *p* is the case. My answer then becomes expressed in an overt utterance or an inner-speech token *p*, which the mindreading system takes as input to interpretively produce the self-ascription *I believe p* as output. The developmental TM account provides a simpler, more computationally plausible proposal on what happens in the context of the transparency of belief.

To see this, note first that even though the account holds that one's initial acceptance of [BSA] depends on various inferences (e.g., pertaining to [ACTBEL] and to one's insight into the basis of one's own acting), these inferences only need to be performed once so as to initiate the re-conceptualization of that what is the case as that what is the case from one's own perspective. This re-conceptualization puts principle [BSA] into place and sanctions future direct transitions from *p* to *I believe p*. Why should we assume that this is so?

The reason pertains to the more general point that the human cognitive system usually tries to operate as economical as it can. There is extensive empirical evidence suggesting that the human mind is a "cognitive miser" in that it tends to short-circuit computations if it can and opt for easy methods (heuristics) rather than difficult ones to solve cognitive tasks (De Neys, Rossi, & Houdé, 2013: 269f; Fiske & Taylor, 2013: 15). Studies suggest that people have a "strong bias to default to the simplest cognitive mechanism" so as to keep processing costs low and preserve limited cognitive resources (e.g., attention; Stanovich, 2011: 30).

If this is right, then the judgment-forming system responsible for producing belief self-ascriptions is likely to adopt simplifying strategies to perform its function too, and, once the re-conceptualization of *p* as *p* from one's own perspective has occurred, implements a direct transition from *p* to *I believe p*. The appeal to simplicity here would not much support the assumption of a direct transition, if the latter relied on processes whose existence is not independently supported (e.g., by their involvement in other tasks than the production of belief self-ascriptions). But according to the developmental TM account, the existence of the processes at issue is independently supported. The transition is simply based on facts to do with one's own agency (see section 4.2) and on a recycling of resources already used to produce other-ascriptions of beliefs (see section 4.1). Since no extra or special processes are postulated, there is

ground to hold that the direct transition captured in [BSA] will be adopted, because this would be more consistent with what we know about the computational constraints under which the human mind operates than the ISA view of the transparency of beliefs. The motivation for assuming that we are in fact likely to often use TM to determine our own beliefs rather than the interpretive processing that the ISA theory proposes is hence that this is better in line with the empirical research that suggests that the human cognitive system is a cognitive miser.

Notice that the claim here is only that the developmental TM account is preferable to the ISA theory when it comes to self-ascriptions of beliefs that figure in conscious first-order order conscious thinking (e.g., on whether *p* is the case).[14] It is consistent with the account that we frequently come to know our other beliefs and PAs via a process of self-interpretation.

5.2 | Objection 2

Goldman (2006: 241f) and Carruthers (2011: 82f) argue that TM accounts of belief self-ascriptions presuppose what they need to explain. On these views, one determines whether one believes, say, that whales are mammals, by asking oneself "Are whales mammals?" The objection then continues that when the response is a judgment that whales are mammals, then "in order for this act of recognition [that whales are mammals] to provide the input necessary" for the mechanism producing belief self-ascriptions, the "information that a *judgment* is occurring with the content, *whales are mammals*, would have to be accessible to whatever mental faculty is charged with applying the rule, '*p*, so I believe that *p*'" (Carruthers, 2011: 82). For the content, *whales are mammals* might be, say, supposed or doubted, rather than judged. Hence, on TM views of belief self-ascriptions, "you would at least need to know your occurrent judgments about what is the case," which is to say that a self-ascription of a judgment must already have occurred for these proposals to work (Carruthers, 2011: 84; Goldman, 2006: 241, Goldman, 2012: 419).

However, the point rests on a misunderstanding. According to TM accounts of belief self-ascriptions such as the developmental proposal introduced above, *S*'s transition from *p* to *I believe p* does not require her to have an insight into the attitudinal component of the representation about *p* any more than a transition in first-order reasoning from *whales are mammals* to the first-order judgment *whales lactate* does. To infer that whales lactate from *whales are mammals*, one *does* need to judge that whales are mammals, rather than suppose or doubt it, but for that to happen, one need not self-ascribe any PA. If, in order to form the judgment that whales lactate on the basis of *whales are mammals*, one had to self-ascribe the attitudes that feed into one's settling whether whales lactate (in addition to the contents) then, by the same line of thought, infants who still lack the ability to represent PAs should also lack the ability to form first-order judgments via simple inferences from other beliefs that they already hold, which is highly implausible. It is also at odds with data showing intact first-order reasoning in individuals with autism who have a deficit in representing PAs (see Scott & Baron-Cohen, 1996; Scott, Baron-Cohen, & Leslie, 1999).[15] These considerations suggest that when *S* judges that *p*, the representation about *p* can play its distinctive functional role as a judgment without *S* representing that functional role itself (see also Byrne, 2012; Peacocke, 1996: 128f). If so, then there is no reason to assume that one first needs to self-ascribe a judgment in order to infer *I believe p* from *p*; one only needs to judge that *p*.

6 | FROM BELIEF SELF-ASCRIPTIONS TO NON-INFERENTIAL SELF-KNOWLEDGE

As it stands, the developmental TM account introduced in the preceding sections is about belief self-ascriptions. These are second-order beliefs that ascribe beliefs to oneself. They do not yet amount to self-knowledge, for beliefs alone do not yet amount to knowledge (Ichikawa & Steup, 2017). A further argument is needed to support the view that the belief self-ascriptions that the TM account pertains to qualify as self-knowledge. I shall now provide one. Moreover, I shall contend that these ascriptions amount to self-knowledge that is both psychologically and epistemically non-inferential in nature.

In order for a belief to qualify as knowledge, the belief needs to satisfy the conditions for knowledge. One plausible condition among them is *safety*. According to Sosa (1999) and Williamson (2001), a belief that *p* is *safe* just in case the belief could not easily have been false. While safety is typically viewed as a necessary condition for knowledge, Byrne (2005: 96) notes that "absent countervailing considerations […], safety can be used as a rough-and-ready diagnostic tool for the presence of knowledge where the proposition in question is contingent." I agree.

With safety as a "diagnostic tool" at hand, it is not difficult to see that belief self-ascriptions formed in the way the developmental TM account suggests qualify as knowledge. This is because whenever *S* self-ascribes a belief on the basis of *p*, she must first have judged that *p*; otherwise, *p* will not be the case for her, from her perspective. Not all judgments issue into beliefs with the same content, as, for instance, non-doxastic factors (biases etc.) may affect them (Peacocke, 1998: 88f). This is why, for instance, Byrne (2005: 96) is arguably mistaken in claiming that the transition from *p* to the self-ascription *I believe p* is always "self-verifying" in that the "resulting second-order belief is true": judgments do not necessarily result in beliefs and the transition at issue is hence not always self-verifying. Still, judgments do reliably give rise to and correlate with beliefs (Cassam, 2010: 83). Furthermore, frequently, when upon wondering whether *p* is the case, one recalls that *p* is the case, then that will be a judgment too. It will be an episode of bringing to consciousness a pre-existing belief (Crane, 2013: 170). In these cases, the correlation between a judgment and the corresponding belief clearly holds. The judgment that *p* thus reliably correlates with the truth-maker of the self-ascription *I believe p*. Since that is so, belief self-ascriptions formed on the basis that *p* is the case cannot easily be false. They are hence safe and can plausibly be viewed as knowledge.

Is this knowledge inferential or non-inferential? To tackle the question, it is useful to follow Cassam (2017: 726f) in drawing a distinction between "psychologically" and "epistemically non-inferential" self-knowledge. Self-knowledge of a belief that *p* is *psychologically* non-inferential if its acquisition does not involve any kind of inference[16] from evidence pertaining to mental states of affairs. It is *epistemically* non-inferential if the "justification for my second-order belief [*I believe p*] doesn't come from my justification for believing any other proposition" (Cassam, 2017: 733). Cassam goes on to argue that "[the transition from *p* to *I believe p*, i.e., TM] itself only delivers inferential knowledge" (Cassam, 2017: 735, footnote 17).

I disagree. Note first that whether or not TM delivers epistemically non-inferential knowledge depends on one's preferred view of epistemic justification. There are roughly two kinds of views on epistemic justification: *externalism* and *internalism* (for discussion, see Goldman, 2009; Kornblith, 2001). According to *externalist* accounts, *S*'s self-ascription *I believe p* might be epistemically justified even in cases when she is not or cannot be aware of the ascription's justifier(s). For instance, as long as the self-ascription *I believe p* is the product of a reliable process, it may still count as justified. In contrast, according to *internalist* accounts, for *S*'s self-ascription *I believe p* to be justified, *S* must be aware of or at least capable of being aware of the ascription's justifier(s), that is, the fact(s) according to which she is justified in self-ascribing the belief that *p*.

While there is much debate on whether externalism or internalism is the correct view of justification (Bonjour & Sosa, 2003; Goldman, 2009; Kornblith, 2001), I shall assume, and add to the developmental TM account, an externalist view of epistemic justification that is in line with safety. I shall take it that a belief that *p* is justified if it is formed via a reliable mechanism, a mechanism that tends to produce true rather than false beliefs.

As it happens, when one concludes that *p* is the case, one's transition from *p* to the self-ascription *I believe p* is a reliable mechanism for belief formation because it involves forming the self-ascription on the basis of the judgment that *p*, which reliably correlates with the belief *p*. The self-ascription is thus externalistically justified. Notice that even if *all* of one's other judgments and beliefs lacked any justification, as long as the self-ascription of the belief that *p* is formed on the basis of the judgment that *p*, and the judgment tends to correlate with the belief that *p*—which it does, even if it is evidentially unsupported—the self-ascription will still be justified. Since that is so, the self-knowledge that the developmental TM account pertains to is epistemically non-inferential in Cassam's (2017) terms.

Is it *psychologically* non-inferential too? It is if it is acquired without any inference from evidence pertaining to mental states of affairs. To assess whether that is so, note first that even if inferences from such evidence are involved

when a particular mechanism for the transition between thoughts is put into place, it does not follow that subsequently the mechanism itself operates in a psychologically inferential way.

According to the TM account introduced above, during cognitive development, a subject *S* links *p* and the self-ascription *I believe p* via a particular set of inferences that do involve evidence pertaining to beliefs. But once the link is established, the mediating inferences are short-circuited to the direct transition from *p* to *I believe p*. When this has happened, upon concluding that *p* at the personal level, *S* does not represent *p* as *p for her*. All that she represents is that *p* is the case, which is hardly evidence for her that she (or anyone) believes anything because *p* might be the case even if no one believes it (Barnett, 2016). Nonetheless, when *S* reflects on whether she believes that *p* and concludes that *p*, then in that situation *p* serves the mindreading system as input to a purely causally operating detection mechanism (e.g., Nichols and Stich's MM)[17] that then issues the self-ascription *I believe p*. This mechanism is put into place and sustained by the inferences outlined above, but those inferences do not figure in its operation. Furthermore, while much of the preceding discussion on the inferences has been pitched at the conscious, personal level, there is no reason to assume that they ever need to enter *S*'s consciousness, for it is well known that the system responsible for mental state ascriptions, i.e., the mindreading faculty, operates mostly unconsciously. There is no reason to deny that the inferences that set up the mechanism that implements TM can be unconsciously executed by that system also. If so, subjects may not have an insight into the basis of their transition from *p* to *I believe p* and may readily grant that *p* is poor evidence for the self-ascription but still treat the transition as sanctioned for reasons that only become manifest to them upon reflection along the lines outlined above.[18] Hence, since the belief self-ascriptions that are formed in the way the developmental TM account proposes do not involve the subject's taking *p* as evidence for some mental state of affairs but are the result of a purely causally operating mechanism, put into place by an unconsciously operating system, these self-ascriptions constitute *psychologically* non-inferential self-knowledge also. Cassam's (2017: 735, footnote 17) claim that "[the transition from *p* to *I believe p*] itself only delivers inferential knowledge" thus needs to be revised, for the developmental TM account indicates that it leads to both epistemically and psychologically non-inferential self-knowledge.

7 | CONCLUSION

I argued that the two most common empirically motivated versions of the asymmetry view of self-knowledge of beliefs, namely Nichols and Stich's (2003) MM theory, and Goldman's (2006) introspection account, have two shortcomings. They (a) leave unexplained how people connect a mental representation about *p* with the self-ascription *I believe p* and (b) are inconsistent with evidence suggesting a self-other symmetry in the development of belief ascriptions.

I then looked at the empirically best-supported version of the symmetry view of self-knowledge of beliefs, the ISA theory, and noted that neither (a) nor (b) are problematic for it. However, the ISA theory holds that we always only have interpretive self-knowledge of beliefs. I maintained that this is too strong, because the empirical data and core arguments supporting the ISA theory do not undermine the developmental TM account introduced, which is a hybrid of both the asymmetry view and the symmetry view and holds that we can and often do acquire non-inferential knowledge of whether we believe that *p* by settling whether *p* is the case.

We can acquire self-knowledge in this way once we apply to ourselves resources that we already rely on to determine other people's beliefs. One of them was the insight that to determine whether a subject *S* believes that *p*, we can simply ask her whether *p* is the case. The other was the thought that subjects do not act directly on the basis of what is the case but on the basis of what they believe to be the case. I argued that this thought, combined with what we know about the basis of our own actions, leads during our cognitive development to a tension that prompts us to re-conceptualize that what is the case as that what we believe to be the case. This re-conceptualization provides a response to the problem captured in (a). Furthermore, since the transition from *p* to the self-ascription *I believe p* is on the account introduced the result of applying to oneself resources already required for other-ascriptions of beliefs,

the belief self-ascriptions at issue should be developmentally posterior to or simultaneous with other-ascriptions of beliefs. Hence, the data relevant for (b) pose no problem for the account either.

Moreover, the account provides an explanation of the transparency of belief that better accommodates the empirical evidence on the operational constraints of the human cognitive system than the one that the ISA theory offers. And it motivates the view that TM is not only frequently used to form belief self-ascriptions but also results in non-inferential self-knowledge. The proposal developed here thus improves on some of the most common empirically motivated versions of both the asymmetry view and the symmetry view of self-knowledge of beliefs, and occupies a space between these two views that has not yet been explored but deserves to be taken seriously.

ACKNOWLEDGMENTS

The account of self-knowledge of beliefs that I introduce here gradually emerged from other similar views I proposed. In this process, I benefitted from discussions with Nick Shea, Tim Bayne, and Tim Crane.

ENDNOTES

[1] Advocates of the asymmetry view typically acknowledge that we sometimes do rely on self-observation or interpretation to come to know our own beliefs (see, e.g., Goldman, 2006; Nichols & Stich, 2003).

[2] Advocates of the symmetry view about beliefs typically grant that there is an asymmetry when it comes to one's access and self-knowledge of other mental states (e.g., sensory-imagistic and affective states; for discussion, see Schwitzgebel, 2014). The symmetry view that I focus on here only pertains to beliefs. That is, it is about a restricted kind of self–other parity.

[3] In this paper, I take the expressions "*p* is the case," "*p* is true," and "*p*" to be interchangeable. I prefer to use "*p* is the case" for stylistic reasons. So on my view here, judging or believing that *p* is the case does not require one to have the concept of something's being the case or the concept of truth in addition to the concept of *p*.

[4] In the cognitive sciences, the term "Box" refers to the functional role type of a mental representation. The term is used in block diagrams, which are common in the description of the architecture of the mind. See Horgan and Tienson (1996: 15f) for details.

[5] For more evidence supporting a developmental symmetry, see Carruthers (2011), Musholt (2012), and Rakoczy (2010).

[6] One often-mentioned proposal is Shallice's (1988) view that PA self-ascriptions might have evolved for executive functions. But see Carruthers (2011: 67) for a response.

[7] See also Happé (2003: 141).

[8] Carruthers does not offer an explanation of *why* the mindreading system was turned inward to begin with. For a plausible proposal, see Shea et al. (2014).

[9] On the basis of data from empirical research including fMRI studies, Carruthers (2011, 2014, 2015) argues that working memory, and so conscious thinking, is sensory-based in that only sensory-imagistic (e.g., mental imagery such as inner speech) and affective representations (e.g., felt desires) can enter working memory but not (non-sensory/non-affective) PAs such as beliefs. They can be accessed in working-memory processing and can figure in conscious thinking. But Carruthers holds that this happens only *indirectly* via sensory-imagistic/affective representations that are broadcast in working memory. The case for this indirect-access view is part of Carruthers' argument for the ISA theory. However, elsewhere I argue that the indirect-access view is problematic (Peters, 2017a); see also Wu (2014). I shall thus set aside Carruthers' working-memory related point for the ISA theory here.

[10] On Carruthers' view, we *do* have non-interpretive access to some of our own attitudes, namely sensorily-embedded judgments, felt desires, and the attitudes that are within the mindreading system's own database (see Carruthers, 2011: 53f). The beliefs that I am concerned with here do not fall into these categories.

[11] There are other accounts of self-knowledge of beliefs that appeal to the transparency of beliefs but do not assume a direct transition from *p* to *I believe p* (see, e.g., Boyle, 2011; Cassam, 2014; Fernández, 2013; Silins, 2012).

[12] In studies using *non-verbal* FBTs that involve measuring the time children spend looking at a particular event (they are thought to look longer at unexpected events), already 7- to 15-month-olds have been found to display some understanding of mental states, as they tend to look longer when Sally goes to B2, where the marble in fact is. This is typically taken to suggest that they already have an *implicit*, that is, an unconscious, procedural, non-conceptual, and automatic grasp of false beliefs (Apperly & Butterfill, 2009; Schneider, Slaughter, & Dux, 2015). My focus here is only on children's *explicit* understanding of beliefs, that is, a conscious, declarative, conceptual, and subject-controlled understanding of beliefs (Low & Perner, 2012).

[13] [BSA] is related to Byrne's (2005: 95) "BEL rule," but there are differences between his view of self-knowledge and mine. I shall come back to this issue below. See also Peters (2017b).

[14] Carruthers (2011, 2014, 2015) argues that PAs such as beliefs are not directly accessible in conscious first-order thinking, because such thinking depends for him on working memory and only sensory-imagistic/affective representations, which are not PAs such as beliefs themselves, can enter this workspace and be directly accessible. If Carruthers' argument for this view were successful, then the proposal I introduced here would be in trouble. However, there is good reason to believe that it is not successful (Wu, 2014), and elsewhere (Peters, 2017a) I argue that the picture of conscious first-order thinking that Carruthers defends faces a number of problems.

[15] See also Peters (2017a).

[16] I take an inference to be a "non-accidental transition between belief contents, where the reasonableness of the transition is open to assessment" (Boyle, 2011: 227). There has been an interesting debate on what exactly an inference is; see Boghossian (2014) and Wright (2014). I shall not delve into the issue here.

[17] It is typically accepted that the outputs of such a mechanism are non-inferentially formed. For instance, Cassam (2010: 91) writes that "Stich and Nichols do not draw attention to the consequences of their view for the issue of immediacy, but it seems obvious that if a Monitoring Mechanism is sufficiently reliable to produce knowledge of one's own beliefs then the knowledge to which it gives rise is both psychologically and epistemically immediate," i.e., non-inferential.

[18] Because subjects *can* make the reasoning explicit, the transition can be intelligible from their own point of view. I develop this point further in Peters (2017b).

ORCID

Uwe Peters http://orcid.org/0000-0002-7103-3921

REFERENCES

Apperly, I., & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970.

Armstrong, D. (1999). *The Mind-Body Problem*. Boulder: Westview Press.

Barnett, D. (2016). Inferential justification and the transparency of belief. *Nous*, *50*(1), 184–212.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1), 1–18.

Bonjour, L., & Sosa, E. (2003). *Epistemic Justification: Internalism vs. Externalism, Foundations vs. Virtues*. Oxford: Blackwell.

Boyle, M. (2011). Transparent self-knowledge. *Proceedings of the Aristotelian Society*, *85*(1), 223–241.

Brasil-Neto, J., Cohen, L., Panizza, M., Nilsson, J., Roth, B., & Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, *55*, 964–966.

Byrne, A. (2005). Introspection. *Philosophical Topics*, *33*(1), 79–104.

Byrne, A. (2011). Transparency, belief, intention. *Aristotelian Society Supplementary*, *85*(1), 201–221.

Byrne, A. (2012). Review of Peter Carruther's The opacity of mind. *Notre Dame Philosophical Reviews*. URL: http://ndpr.nd.edu/news/the-opacity-of-mind-an-integrative-theory-of-self-knowledge/

Byrne, R., & Whiten, A. (1997). *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.

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

Carruthers, P. (2014). On central cognition. *Philosophical Studies*, *170*(1), 143–162.

Carruthers, P. (2015). *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*. Oxford: Oxford University Press.

Cassam, Q. (2010). Judging, believing and thinking. *Philosophical Issues*, *20*, 80–95.

Cassam, Q. (2014). *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Cassam, Q. (2017). What asymmetry? Knowledge of self, knowledge of others, and the inferentialist challeng. *Synthese*, *194*, 723–741.

Crane, T. (2013). Unconscious belief and conscious thought. In U. Kriegel (Ed.), *Phenomenal Intentionality* (pp. 156–173). Oxford: Oxford University Press.

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*, 568–570.

Dennett, D. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole, & D. Johnson (Eds.), *Self and Consciousness: Multiple Perspectives* (pp. 103–115). Hillsdale, NJ: Erlbaum.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin and Review*, *20*(2), 269–273.

Dretske, F. (2003). Externalism and Self Knowledge. In S. Nuccetelli (Ed.), *Semanatic Externalism, Skepticism and Self-Knowledge* (pp. 131–143). Cambridge, MA: MIT Press.

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

Fernández, J. (2013). *Transparent Minds: A Study of Self-Knowledge*. Oxford: Oxford University Press.

Fiske, S., & Taylor, S. (2013). *Social cognition*. London: Sage.

Gallois, A. (1996). *The World Without, the Mind Within: An Essay on First-Personal Authority*. Cambridge: Cambridge University Press.

Goldman, A. (2006). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Goldman, A. (2009). Internalism, externalism, and the architecture of justification. *Journal of Philosophy*, *106*(6), 309–338.

Goldman, A. (2012). Theory of mind. In E. Margolis, R. Samuels, & S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 402–424). Oxford: Oxford University Press.

Gopnik, A. (1993). The illusion of first-person knowledge of intentionality. *Behavioural and Brain Sciences*, *16*, 1–14.

Gordon, R. (2007). Ascent routines for propositional attitudes. *Synthese*, *159*(2), 151–165.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, *117*(1), 54–61.

Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences*, *1001*, 134–144.

Harris, P. (1989). *Children and Emotion: The Development of Psychological Understanding*. Oxford: Blackwell.

Horgan, T., & Tienson, J. (1996). *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.

Hornsby, J. (2008). A disjunctive conception of acting for reasons. In A. Haddock & F. Macpherson (Eds.), *Disjunctivism*. Oxford: OUP.

Ichikawa, J., & Steup, M. (2017). The Analysis of Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.). Stanford, USA: The Metaphysics Research Lab Center for the Study of Language and Information, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/knowledge-analysis/

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

Kolodny, N. (2005). Why be rational?. *Mind*, *114*, 509–563.

Kornblith, H. (2001). *Epistemology: Internalism and Externalism*. Oxford: Blackwell Publishers.

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, *30*(1), 1–13.

Moran, R. (2001). *Authority and Estrangement*. Princeton: Princeton University Press.

Musholt, K. (2012). Self-consciousness and intersubjectivity. *Grazer Philosophische Studien*, *84*, 63–89.

Nichols, S., & Stich, S. (2003). *Mindreading*. New York: Oxford University Press.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Peacocke, C. (1996). Our entitlement to self-knowledge: Entitlement, self-knowledge, and conceptual redeployment. *Proceedings of the Aristotelian Society*, *96*, 117–158.

Peacocke, C. (1998). Conscious Attitudes, Attention, and Self-Knowledge. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing Our Own Minds* (pp. 63–98). Oxford: Clarendon Press.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences*, *16*(10), 519–525.

Peters, U. (2017a). On the accessibility of beliefs in conscious thinking. Manuscript in preparation.

Peters, U. (2017b). The intelligibility of the transparency of belief. Manuscript in preparation.

Rakoczy, H. (2010). Executive function and the development of belief-desire psychology. *Developmental Science*, *13*(4), 648–661.

Robbins, P. (2004). Knowing me, knowing you: Theory of mind and the machinery of introspection. *Journal of Consciousness Studies*, *11*(7/8), 129–143.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Schneider, D., Slaughter, V., & Dux, P. (2015). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*, *22*(1), 1–12.

Schwitzgebel, E. (2014). Introspection. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Stanford, USA: The Metaphysics Research Lab Center for the Study of Language and Information, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/introspection/

Scott, F., & Baron-Cohen, S. (1996). Logical, analogical, and psychological reasoning in autism: A test of the Cosmides theory. *Development and Psychopathology*, *8*, 235–246.

Scott, F., Baron-Cohen, S., & Leslie, A. (1999). 'If pigs could fly': A test of counterfactual reasoning and pretence in children with autism. *British Journal of Developmental Psychology*, *17*, 349–362.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Silins, N. (2012). Judgment as a guide to belief. In D. Smithies, & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 295–328). Oxford: Oxford University Press.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–153.

Stanovich, K. (2011). *Rationality and the Reflective Mind*. New York, NY: Oxford University Press.

Stephens, G., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge, MA: MIT Press.

Sterelny, K. (2012). *The Evolved Apprentice: how evolution made humans unique*. Cambridge, MA: MIT Press.

Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press.

Velleman, D. (2006). *Self to self: Selected essays*. Cambridge: Cambridge University Press.

Wegner, D., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*, 480–491.

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind-development: The truth about false belief. *Child Development*, *72*, 665–684.

Williams, D., & Happé, F. (2010). Representing intentions in self and other: Studies of autism and typical development. *Developmental Science*, *13*(2), 307–319.

Williamson, T. (2001). *Knowledge and its limits*. Oxford: Oxford University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Wright, C. (2014). Comment on Paul Boghossian, 'what is inference'. *Philosophical Studies*, *169*(1), 27–37.

Wu, W. (2014). Being in the workspace, from a neural point of view: Comments on peter carruthers, 'On Central Cognition'. *Philosophical Studies*, *170*(1), 163–174.

How to cite this article: Peters U. Introspection, mindreading, and the transparency of belief. *Eur J Philos*. 2018;26:1086–1102. https://doi.org/10.1111/ejop.12318