Transparency Over-Extended[*]

Annalisa Coliva
(University of California, Irvine)

Edward Mark
(University of California, Irvine)

Abstract: In this paper, we argue that epistemic accounts of transparency of the sort put forward by Alex Byrne (2018) and Jordi Fernández (2013) cannot offer a sufficient explanation of the first-personal knowledge we have of our own mental states. We begin by rehearsing Paul Boghossian's trilemma about self-knowledge (**§1**). We then identify how the two aforementioned accounts purport to be ways out of this trilemma. We argue against the plausibility of their strategy by noticing that these accounts either (i) fail to present an epistemic account; (ii) assume the very knowledge they are designed to explain (i.e. knowledge of one's first-order mental states); or, (iii) endorse a dubious inferentialist story of how we move from being in a given first-order mental state to its knowledgeable self-ascription (**§2-3**). Finally, we close by highlighting the difficulties involved in presenting these accounts as explanatory for states other than belief (**§4**) and move to suggest a pluralist approach to the study of self-knowledge as a means for moving beyond Boghossian's trilemma (**§5**).

Introduction

The knowledge that we have of at least some of our mental states appears to be marked by several distinctive features. The beliefs we have about our own mental states seem more likely to amount to knowledge than our beliefs about, say, objects or other minds. In this sense, we can say

[*] Coliva is the main author of §§2.i, 2.ii, 3.i, 3.ii, Mark is the main author of §§1, 4, 4.i, 4.ii, the remaining sections have been elaborated together.

that self-knowledge is 'privileged'. Additionally, the method by means of which we come to know our own mental states seems to be one that is not available to others in the same way it is available to us. In this sense, self-knowledge is 'peculiar'.[1]

One way to explain these features of self-knowledge is to claim that self-knowledge is 'transparent'. In *The Varieties of Reference*, Gareth Evans famously suggests that "in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward-upon the world." He then goes on to observe "If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?'" (Evans 1982: 225). This passage has served as a starting point for a number of theorists who attempt to explain the knowledge we have of our own mental states by arguing that we look, not to the mental state itself, but to some relevant state of affairs in the world. These 'transparency accounts' as they have come to be called take a variety of forms. For agentialists like Tyler Burge (1996) and Richard Moran (2001), the capacity for self-knowledge is grounded in the nature of rational agency. Epistemic transparency accounts proposed by Alex Byrne (2018) and Jordi Fernández (2013) attempt to provide an explanation of self-knowledge by reference to an empirical justification for the self-attribution of the first-order state. This justification for the self-attribution makes no reference to the mental states themselves or to any ability to introspect such states. It is this epistemic version of transparency with which we will be concerned in this essay.

These epistemic accounts of self-knowledge promise a number of advantages. First, they seem to provide an explanation for why self-knowledge is both privileged and peculiar. Second, they purport to be economical. That is, they claim to offer an explanation of self-knowledge that relies on familiar

[1] Unfortunately, a standardized terminology has yet to be accepted here. In using 'privileged' and 'peculiar' to denote these features of self-knowledge, we have adopted Byrne's terminology; however, some authors utilize a different set of terms to pick out these same features. Fernández will use 'strong' and 'special' to track a similar distinction.

epistemic capacities that are needed for knowledge of other subject matters. Finally, these accounts can be deployed, with little to no modification, to explain the knowledge we have of a wide variety of mental states. Given all of this, epistemic transparency accounts can be quite appealing; however, we will argue that these accounts cannot make good on their promise.

The plan is as follows: we begin by rehearsing Boghossian's now classic trilemma about self-knowledge: self-knowledge is either based on observation, on inference, or on nothing (§**1**). We then briefly note the way in which epistemic accounts purport to provide a solution to this trilemma (§**2-3**). We argue against the success of their strategy by showing that these accounts either (i) fail to present an epistemic explanation of self-knowledge, (ii) assume the very knowledge they are designed to explain - i.e. knowledge of one's first-order mental states, or (iii) endorse a dubious inferentialist story regarding how we move from being in a given first-order mental state to its knowledgeable self-ascription. We will close by highlighting the difficulties inherent in extending epistemic transparency accounts of self-knowledge to mental states other than belief (§**4**), and suggest a pluralist approach to the study of self-knowledge as a means for moving beyond Boghossian's trilemma.

§1. Boghossian's Trilemma

In "Content and Self-knowledge," Boghossian observes that there are three possible ways in which one might explain the knowledge we have of our own mental states: (i) by inference, (ii) by introspection, or (iii) by nothing at all. Boghossian argues that all three explanations are flawed. Though not one of the available explanations is sufficient, our capacity for self-knowledge "is not an optional component of our ordinary self-conception" (Boghossian 1989: 6).

That we might know our thoughts by inference from other beliefs is quickly dismissed by Boghossian. He notes that on a straightforward construal of the inferential model, we must believe both that the relevant evidence obtains and that this evidence justifies the target proposition. However, if this model is applied to self-knowledge, a regress ensues. By hypothesis, it is stipulated that the second-order belief is justified by an inference from the first-order belief. However, if the first-order belief is to serve as evidence for the second-order belief, we must already have knowledge of the first-order belief. Thus, an inferential account of self-knowledge will assume exactly what it sets out to prove insofar as we must appeal to an instance of self-knowledge in order to explain our justification for the second-order belief.

The idea that we know our thoughts through introspection does not fare any better. Boghossian notes that, if introspection is understood in a way that is analogous to perception, then introspection would only make manifest intrinsic features of our mental states, not their extrinsic features. Accordingly, he concludes that introspection cannot account for self-knowledge. Boghossian's argument involves two distinct premises. First, Boghossian assumes that perception provides information regarding intrinsic, non-relational properties. This information might be used to infer relational properties, but this is not something that would be given in perception itself.[2] For example, the content of our visual experience provides information regarding the color and shape of a coffee cup in front of us. However, the content of this experience would not include information regarding the fact that this cup is one of a set of three (let us stipulate that the remaining members of this set are currently not within our field of vision). This kind of relational property is simply not given by perceptual content. Second, Boghossian assumes a version of content externalism according to which the content of (at least some) of one's mental states is at least partially individuated by extrinsic features.[3] From these two premises, Boghossian concludes that we cannot know all of our mental

[2] This model of perception is more or less questionable depending on how rich or poor one takes perceptual content to be. However, our target is not the viability of Boghossian's treatment of perception so we won't press the issue further.

[3] To be clear, the claim here is relatively modest. It is not the case that all of our mental states are so individuated. Rather, it is enough for at least some mental states that can be known to be individuated in this way. For more on this see, for example, (McKinsey 1991: 9-16) or (Boghossian 1997: 161-75).

states by introspection after all. Though knowledge of some of these extrinsic features is needed in order for one to have knowledge of the mental state, the relation between the content of our mental state and environmental conditions is not something that would be made available through introspection alone.

Finally, Boghossian claims that it is, at best, unlikely that we have knowledge of our own minds based on nothing at all. He notes that if self-knowledge were based on nothing at all, then the account would be unable to explain a set of features commonly associated with self-knowledge; namely, that self-knowledge is (i) directed, (ii) gradated, and (iii) fallible. The claim that we might be fallible regarding our mental states is straightforwardly true—at least for some set of mental states. To claim that self-knowledge is directed is to claim that we are able to decide how much attention to pay to our mental states. This is closely related to the claim that our self-knowledge is gradated—i.e., that some individuals are better than others at reporting on their mental states. Boghossian argues that any sufficiently developed theory of self-knowledge should be able to explain these phenomena; yet, were one to argue that knowledge of our own mental states is based on nothing at all, this would provide no insight as to how and why our self-knowledge exhibits these features.

None of the three explanations Boghossian considers offer a satisfactory account of self-knowledge. Insofar as self-knowledge is an essential aspect of our theory of mind, we are thus forced to choose between three equally bad options. At least, this is Boghossian's claim. While there is certainly room to question whether the trilemma he offers is genuine, it is clear that any convincing account of self-knowledge must either (i) provide an argument for why one of the three options considered by Boghossian remains viable, (ii) offer an alternative epistemic explanation of our self-knowledge that does not lean on any of the three explanations considered by Boghossian, or (iii) be able to provide some explanatory yet non-epistemic account of self-knowledge.[4] In the remainder of this essay, we will argue that responding to Boghossian's challenge turns out to be uniquely difficult for transparency accounts of the epistemic variety.

In the following two sections we will spell out the specifics of two prominent epistemic accounts of transparency; namely, those of Jordi Fernández and Alex Byrne. We will argue that neither account successfully answers Boghossian's challenge. Fernández, in an attempt to avoid the pitfalls associated with an inferentialist or inner-sense account, proves unable to provide an epistemic explanation of how we move from being in a first-order mental state to its knowledgeable self-ascription. Byrne, for his part, defends an inferentialist approach, but in the end, relies on a dubious conception of the first-order state that supposedly serves as grounds for a rational self-ascription of the second-order state. As such, Boghossian's concerns regarding circularity remain very much alive.

§2.i Fernández and the Bypass Model

Transparency theorists are committed to the general claim that one knows one's own mental states by attending not to these mental states themselves, but to the world 'outside' of one's mind. According to what Fernández dubs the 'Bypass Model', we are able to self-ascribe beliefs because we enjoy certain mental states (such as experiences, memories, etc.) which typically lead us to form both the first order belief as well as the self-ascription of that belief. He writes,

> …suppose that the support that I have for believing that I have some belief is identical with my grounds for that belief. Then, it makes sense that, in order to address the question of whether I have a belief about the world, I attend to the world. I am simply looking for adequate support for my self-attribution of the relevant belief. After all, this is what I normally do when

[4] Peacocke's 'reasons account' and various Constitutivist accounts might be seen as examples of the second and third options respectively. In this essay we will not be able to assess the merits of these accounts; rather we simply hope to show that where the 'reasons account' and the Constitutivist might have convincing responses to Boghossian's trilemma, the transparency theorist does not fare quite so well.

> I am about to form any of my beliefs rationally; I look for adequate support for the belief. (Fernández 2013: 49-50).

So, for example, in self-attributing the belief that the coffee is cold, we bypass the first order state and attend directly to the relevant grounds for the first order state (e.g., the perception that the coffee is cold).

Fernández argues that the Bypass Model can explain the privileged and peculiar access requirement insofar as we enjoy certain mental states which naturally and usually lead us to form the corresponding beliefs as well as second-order self-ascriptions, without any need to observe our own overt behavior and without having to reason from the basing mental state to the second-order self-ascription (Fernández 2013: 57). By contrast, other people will be in a position to ascribe a belief to us only by observing our overt behavior and by inferring to its likely cause. If, however, our self-ascriptions aren't based on outer observation and inference to the best explanation, they aren't open to the kinds of error that can beset third-personal ascriptions of belief. The access we have to our own beliefs is different and also more secure than whatever access other people might have to them. Fernández also claims that the Bypass Model isn't causal because the self-ascription isn't causally brought about by the first-order belief, but is instead grounded on the evidence on which that very belief is based.

Two features of this account are worth reiterating: first, the view is intended to be non-inferential. According to Fernández, forming the second-order belief "requires neither reasoning nor behavioral evidence." (Fernández 2013: 56). Recall that any view that claims to be inferential seemingly faces issues of circularity. That is, on a straightforward reading of an inferentialist account, the inference utilized to arrive at the self-ascription seems spurious insofar as it is an inference based on knowledge of the first-order state itself. According to the straightforward reading, it is the knowledge of the first order state itself that would justify the self-ascription of that state insofar as the mere truth of the content of the first order state is poor evidence that one has the corresponding belief. Fernández can avoid this particular issue only if he can provide an explanation for how the second-order state is formed such that its formation (i) reliably tracks the first-order state and (ii) is not based on an inference from knowledge of the relevant first-order state. Second, the view is purportedly non-causal. This is an important feature of the Bypass Model. If the self-ascription were causally entailed by the first order state, this would seem to suggest that self-knowledge is not a cognitive achievement; rather, it would be a state in which we simply seem to find ourselves. Since, in keeping with Boghossian, Fernández takes self-knowledge to be a cognitive achievement, it is important for his account that he is able to provide a non-causal explanation for the self-ascription of the second-order state. So, while Fernández does make reference to causal relations in his account, he takes himself to be doing so in a way that does not preclude him from claiming that self-knowledge is a cognitive achievement. To quote Fernández at length

> The model does not assume that, in order for the self-attribution of a belief to give one privileged access to it, that the self-attribution must have been caused by the self-attributed belief. In that sense, the bypass model is a non-causal account of self-knowledge. (the term 'bypass' is meant to suggest that feature of it.) however, it requires certain kinds of causal relations to be in place. The bypass model assumes that, if a self-attribution of a belief is justified, then two kinds of causal relations are in place. On the one hand, the self-attribution must have been caused by the state that constitutes the subject's grounds for the self-attributed belief…on the other hand, a robust enough correlation must hold between the type of state that constitutes the subject's ground for the self-attributed belief and that belief. And such a correlation rests on a causal relation as well. (Fernández 2013: 61).

Despite this, Fernández will insist that the Bypass Model is not causal—at least not in a way that would jeopardize his claim that self-knowledge is a cognitive achievement. Though, as we will soon see, it is not perfectly clear how this works, the contrast case for Fernández is meant to be a sort of functionalist account of self-knowledge wherein being in one mental state simply causes one to have the relevant second-order state. At least relative to the functionalist picture, the Bypass Model is non-causal and so purportedly will allow for self-knowledge to constitute some kind of cognitive achievement.

§2.ii. Problems with the Bypass Model

The main claim of the Bypass Model is that self-ascriptions of belief should be justified by the evidence that prompts the first-order belief itself. We find this to be a problematic way of justifying knowledgeable self-ascriptions. Intuitively, the self-ascription, if justified at all, should be justified by the corresponding first-order mental state. Indeed, it seems reasonable to think that while we may have unjustified first-order beliefs, if they were formed on inappropriate, scant, or no grounds, we could still be justified in self-ascribing them for the simple reason that we do have them.[5] To rescue this compelling intuition, Fernández has to say that while the grounds for the subject's first-order belief are insufficient for the justification of the first-order belief, those very grounds would count as a justification for her self-ascription (Fernández 2013: 66). Now, it is clear that this account would work only if we considered the role of our evidence in giving rise to the first-order belief which would then tend to correlate with one's self-ascription of it. But, if that's the case, in the end what "justifies" us in self-ascribing the belief is precisely the first-order belief itself, not the evidence that prompts it.

Moreover, the account leans heavily on the fact that once we are exposed to certain kinds of evidence, we tend to form the corresponding first-order beliefs and, if conceptually equipped, we also tend to form the relevant psychological self-ascriptions. The strength of this correlation is what is meant to secure privileged access. Yet, the account is supposed to do this in a way that does not negate the fact that self-knowledge constitutes some kind of cognitive achievement. We find all of this to be a bit opaque. True, if the story Fernández gives were correct, then the causal basis of the psychological self-ascription wouldn't be the first-order belief; but, as Fernández himself admits, the eventual psychological self-ascription would be arrived at by means of a causal process nonetheless. In fact, the account rests on two causal processes: first, between the grounds for the first-order state and the

[5] We're operating under the assumption that evidence or grounds, which we have here used interchangeably, can lend justification to a belief; though, clearly, one could have some evidence or grounds and yet fail to have a justified belief.

second-order state and then again between the grounds for the first-order state and the first-order state itself (Fernández 2013: 61). What is more, these must function in tandem. In the end, this does not seem to allow for self-knowledge to constitute a cognitive achievement any more than the functionalist model does. That Fernández can maintain this conjunction of causes, between the grounds and the second-order state on the one hand and the grounds and the first-order state on the other, all while maintaining that self-knowledge is a cognitive achievement strains credulity. And though we might think that we could resolve this problem by only making reference to the relevant correlations, without some kind of causal relation, it is unclear that Fernández can establish a connection between the first-order state and the second-order state that is sufficiently strong so as to explain the privileged access we have to our own mental states.

So, the Bypass Model struggles on at least two fronts: first, in order to maintain that it is possible for us to justifiably self-ascribe an unjustified first-order belief, it must be the case that the self-ascription is justified not by the grounds for the first order state, but on the first-order state itself. Second, Fernández is not able to maintain that self-knowledge is a cognitive achievement and that it is privileged. Returning to Boghossian's trilemma, we can frame the problem as follows: given that the Bypass Model is, by Fernández's own admission, neither inferentialist nor based on introspection, and given that the 'no-view' option is anathema to an epistemic account of transparency, Fernández can address the trilemma only if he can offer a viable 'fourth way' out. Now, in theory, a causal explanation could offer just this, but we do not see how Fernández can offer a sufficiently robust account based on cause and maintain that self-knowledge is a cognitive achievement. While Fernández has avoided the pitfalls of the first two options, he has not provided a viable alternative. The fourth way has foundered. All of this should give us sufficient reason to explore alternative options.

3.i. Byrne's Inferentialist Model

Contrary to Fernández, Byrne's account is explicitly inferentialist in nature. Byrne claims that it is by following rules such as BEL – 'If *p*, believe that you believe that *p*' – that one acquires knowledge regarding one's beliefs. One can be said to follow BEL if one self-ascribes the belief in *p*, *because* one recognizes that *p* obtains where 'because' indicates a kind of reason-giving causal connection (i.e., a 'basing relation'). It is important to note that, for Byrne, rule-following is not self-intimating. That is, one can follow a rule like BEL without realizing that one is doing so. There is no need for the individual to understand or endorse the connection between the contents of the rule (Byrne 2018: 102). This constitutes a significant departure from Boghossian's conception of the relevant inference. Recall that for Boghossian, for one to infer from some mental state, that mental state must first be known to the reasoner. Byrne, in denying this assumption, is attempting to revitalize the inferentialist option that Boghossian has discounted.[6]

Byrne admits that, at first glance, BEL does not look particularly promising. First, the transition that underlies BEL seems to be neither a deductively valid inference, nor one based on induction or abduction (Byrne 2018: 75-76). Second, reasoning in accordance with BEL does not appear to be reliable, since there are many truths one does not believe (Byrne 2018: 100). Third, the antecedent of BEL does not provide strong evidence for the consequent since the fact that *p*, by itself, is inadequate support for the hypothesis that one believes that *p* (ibid.). Finally, in certain cases, BEL will license

[6] We should note at this point that, for Byrne, it is not necessary that the subject recognize the evidence from which they infer to be evidence as such. It is not a process of 'critical reasoning' wherein the subject is aware of the conditions guiding the process. Byrne's conception of inference (i.e., reasoning) is merely as some causal transition between belief states (Byrne 2018: 15). This is distinct from the way Boghossian cashed out the concept of inference and may address the simple version of the regress; for, Boghossian's concern was that one needed to take one's first-order belief as evidence, something which presumably requires self-knowledge. If Byrne has discarded the requirement that an inference must be based on evidence that is recognized by the subject as playing an evidential role, then the basic formulation of Boghossian's critique of Inferentialism fails to scathe Byrne. However, as we will see, the way that the subject is related to the premise of Byrne's inference is nonetheless fraught.

certain self-ascriptions even if the inference contains a false step; for example, when one believes a false proposition and so self-ascribes the corresponding belief (ibid.).

Nevertheless, Byrne argues that BEL is defensible. The key move consists in claiming that if one *recognizes* that *p* is the case, then one will *ipso facto* form the belief that *p* (Byrne 2018: 104). BEL is thus 'self-verifying' since if one follows BEL, one will form the belief that one believes that *p*. Furthermore, BEL is 'strongly self-verifying' since even if *p* is not the case, if one mistakenly thinks that *p* is the case, one will thereby form the belief that *p*. It seems that by simply trying to follow BEL, one will form the true belief that one believes that *p*. This means one's belief is *safe* even if the evidence for *p* is not reliable. On this assessment, first-personal knowledge of beliefs is quite clearly inferential since the process one goes through by following BEL is a causal transition from believing that *p* to believing that one believes that *p*. In this formulation, the transparency account explains why this knowledge is peculiar, for recognizing that *p* is the case can bring about the belief that *p* only in one's own case. The corresponding third-personal rule – 'If *p*, believe that S believes that *p*' – would not hold. That is to say, it would not generate safe beliefs. Additionally, since BEL is strongly self-verifying, privileged access is explained as well.

3.ii. Problems with Byrne's Inferentialist Model

Despite Byrne's defense, we believe that BEL, and Byrne's inferentialist model more generally, face considerable challenges. In order to see why, it will be helpful to expand a bit on Byrne's notion of rule following. Byrne notes that S follows the rule R ("if condition C obtain, believe that *p*") if and only if (i) S believes that *p* because she recognizes that conditions C obtain; this implies that (ii) S recognizes (hence knows) that conditions C obtain; (iii) conditions C do in fact obtain; (iv) S believes that *p* (Byrne 2018: 101). However, if this is how one is to understand what it means to follow BEL, there seems to be a significant problem with Byrne's account. Supposedly, we follow the rule stated in BEL and actually infer from *p* that we believe that *p*. This would not be mysterious if we went from consciously holding that *p* is the case, thus forming the belief that *p*, to forming the belief that we believe that *p* by consciously following BEL. That is, by recognizing that one holds *p* true and that, if so, one is entitled to conclude that one believes that one believes that *p*. In that case, the inference would be psychologically real and the initial mental state we would be in, together with the fact that we consciously follow BEL, would causally (and rationally) explain our forming the second-order belief. Yet, Byrne does not think we need to be aware of following BEL in order to be following it (Byrne 2018: 102 & 114). If that is true, we are left to wonder what evidence there would be to support the claim that an actual inferential process in which each state we are in, together with the rule for the transition, is causally responsible for the manifestation of the second-order belief. At most, it would be a lucky guess if it were so.[7]

This conclusion is clearly undesirable; however, the available alternative is no more appealing. Indeed, if Byrne opts for saying that we consciously follow the rule stated in BEL, a different, but equally disconcerting issue surfaces. As Matthew Boyle claims, if the basis for the inference were one's representing oneself as accepting that *p* is the case, that would "presuppose that I already know my mind on the matter, and that would undermine Byrne's account" (Boyle 2011: 230). Thus, Byrne's transparency account would provide a circular explanation of self-knowledge. For the very knowledge of one's belief that *p* would be presupposed in order for one to be able to follow the rule stated in BEL. Alternatively, if the starting point of the inference is *p* itself and not one's representing oneself as believing that *p*, then the inference would be "mad" (ibid.). If one asks oneself why one believes

[7] What we have said here does not rule out the possibility that there is a subpersonal inference, it only suggests that we lack any direct evidence to support it. We hasten to note, however, that even if we countenance the idea of a subpersonal rational inference, there would be something odd about this particular inference. As Byrne himself notes, the transition here is neither deductively valid, nor based on induction or abduction, nor reliable, nor will the antecedent of BEL provide strong evidence for the consequent. Once one maintains that it is also subpersonal, it is odd to say the least to maintain that this transition is in fact an inference at all.

that one believes that *p*, "the answer 'P' is obviously irrelevant… a modicum of rational insight will inform me that, even if it is true that *P*, this by itself has no tendency to show that I believe it" (ibid.). Thus, Byrne's account faces a dilemma, for it is either hopelessly circular as Boghossian predicted, or it is such as to impute a mad inference to subjects.[8]

It seems then that both accounts break down when it comes to providing an explanation of the relation between the first-order state and the knowledgeable self-ascription of that state. Byrne's inferential strategy faces issues of circularity, but Fernández's attempt to avoid the inferentialist strategy either collapses into a causal view or fails to provide an epistemic account of self-knowledge.

§4. Extending the Transparency Account

In **§§2-3**, we argued that the two most prominent epistemic transparency accounts, those articulated by Jordi Fernández and Alex Byrne, cannot adequately explain how we come to know our own beliefs. Now, despite the fact that these accounts face significant issues for what is supposed to be their paradigm case (i.e., knowledge of our own beliefs), both Byrne and Fernández attempt to generalize their models. Fernández, for his part, moves to generalize to a small set of propositional states; Byrne, however, thinks that his version of the transparency account generalizes to all mental states about which we can have privileged and peculiar knowledge. Both the specifics of these claims as well as the general motivation for the generalization strike us as problematic.

In what follows, we will argue that the move to extrapolate these models is misguided for two reasons. First, variations on the problems noted in **§2.ii** and **§3.ii** resurface when these accounts are deployed to try to explain self-knowledge of states other than belief. Second, the inclination to find a

[8] Byrne has considered Boyle's objection. See (Byrne 2018: 123-4). However, we remain unconvinced that the objection is answered.

single account that is capable of making sense of self-knowledge in all of its variety is itself unfounded. In the following section will look at the details of both accounts as they pertain to knowledge of desires. We will then move to question the legitimacy of the general assumption that a single account might be offered as an explanation for self-knowledge.

§4.i Fernández and Desire

Fernández reasons that if the Bypass Model is well positioned to explain the privileged and peculiar access we have to our beliefs, it should be equally well suited to explain knowledge of any other state that is both (i) propositional and (ii) accessible to us in this uniquely first-personal way. Though we arguably have privileged and peculiar access to a wide variety of mental states, Fernández focuses on developing the Bypass Model as an explanation for how we come to know our own desires.[9] In keeping with his account of how we come to know our own beliefs, his goal is to show that, in forming second-order beliefs about our desires, we look past the desire itself to the grounds for that desire.

Essential to Fernández's account is the claim that when we self-ascribe a desire *D*, there is some other mental state *G* such that usually, when we occupy G, we have *D* (Fernández 2013: 82). What is more, he claims that the same state, *G*, is the grounds for both the desire and the self-ascription of that desire. So, for example, thirst tends to be associated with having a desire for a drink. The self-ascription of the desire for a drink is justified by reference to thirst. In this example, thirst serves as both the grounds for the desire for a drink and the grounds for the self-ascription of that desire. Thus, Fernández claims that the privileged and peculiar access we have to our own desires is a result of the

[9] Fernández recognizes that there may be desires that we have that are not available to us in this unique way. His claim is not that all of our desires are first-personally available. Rather, the suggestion is that when we do know our own mental states first-personally, we will have arrived at this knowledge in the way he describes.

fact that, in normal circumstances, we form beliefs about our desires on the basis of our grounds for those desires (Fernández 2013: 86). Fernández claims that though neither causally nor inferentially related, there is a correlation between being in the first-order mental state and the self-ascription of that state. Fernández explains this relation in terms of 'tendency laws' which "specify some circumstance in which, usually (but not always) a subject acquires a desire" (Fernández 2013: 84).

However, this cannot be correct. While we might assess the grounds for our desire and even recognize circumstances that usually attend our experience of some particular desire, this is clearly insufficient as a means for accurately self-ascribing that desire. Fernández's example makes this a bit hard to see. He writes that when we know that we desire a drink, the grounds for this is the fact that we are thirsty. Now, 'thirsty' here cannot mean something like 'having a desire for a drink' as this would render the account hopelessly circular. Rather, Fernández understands 'thirst' as a sort of brute urge where an urge is "a state wherein the subject experiences the fact that she is not in some state as unpleasant." (Fernández 2013: 83). So, the fact that we are not satiated leads to both the desire for a drink and the self-ascription of that desire. But, there is an obvious problem with this: though it might feel odd to say both that we are thirsty and that we do not desire a drink, it is clear that we can be in an unsatiated state and, for some reason, fail to want a drink while knowing that this is so. Perhaps, though thirsty, we are also nauseated. In this case we would not desire a drink and we would not self-ascribe the desire. So, the Bypass Model predicts that we would self-ascribe a desire, when in fact we would not. Note that this critique can be run in the other direction as well. It is clearly possible that we can, while not being thirsty, still desire a drink and know that we do. Perhaps we have a bad taste in our mouths or have just had some particularly spicy food. Thus, the Bypass Model, when applied to desires, both over-generates and under-generates the self-ascription of desire.

One response to this might be to say that when these cases occur, they will still be based on some assessment of grounds that we come to the conclusion that we either desire a drink or do not desire a drink. Perhaps there are other factors that we recognize that indicate a countervailing desire.[10] We might redescribe the case of overgeneration and say that, though unsatiated and though we might usually desire a drink under these circumstances, the fact that we are nauseated is grounds for the self-ascription of the desire *not* to have a drink. Similarly for the case of under-generation, we might claim that while not being thirsty, we still desire a drink and know that we do because we have recognized some additional grounds for the desire (e.g., we have a bad taste in our mouth). Could the Bypass Model be developed in this way? We think not. The account would have to maintain that knowledge of our desires is the result of a complex sort of deliberation in which we weigh conflicting evidence and come to a conclusion regarding what we desire. Apart from the fact that this does not seem to accord with what it is like to know our desires in most cases, it is unclear how to specify this process in a way that avoids reference to any causal or inferential connection between the grounds and the self-ascription.

The examples of over-generation and under-generation are symptoms of a larger issue. As we have argued regarding belief, in the case of desire, the Bypass Model cannot provide a sufficient link between the first-order state and the knowledgeable self-ascription of the second-order state. The 'tendency laws' that Fernández appeals to, as merely specifying circumstance in which usually a subject acquires a desire, are insufficient to the task. What is more, there appears to be no substantive alternative available to Fernández. As we have seen, taking an inferentialist tack is one way to establish the appropriate link between the first-order state and the second-order state such that privileged and peculiar access can be explained; however, cautious of the threat of circularity, Fernández rejects this option. Another possible way to secure the second-order state to the first-order state without involving inference would be to argue that the two states are somehow causally linked. Yet, as noted, Fernández

[10] Fernández is careful to note that an urge is different from a desire to occupy the state the absence of which is experienced as an urge (Fernández 2013, p. 83).

is compelled to avoid this if he is to maintain that self-knowledge is a cognitive achievement. We are, once again, left wondering what exactly links the knowledgeable self-ascription to the first-order desire such that the account can explain our privileged and peculiar access to our desires. In the absence of an answer to this, the account does not offer an obviously epistemic explanation of our self-knowledge.

§4.ii Byrne and Desire

We have seen that the Bypass Model has difficulty tying the relevant basing state to the self-ascription of desire. As it happens, Byrne's inferentialist account does not fare much better. Though Byrne will deploy his model as a means for explaining the self-knowledge we have regarding a wide variety of mental states, for the sake of continuity, we will focus on Byrne's account of desire.

Byrne takes it that a scheme much like BEL can be established for desire. In this case it is by following DES – 'If *φ*ing is a desirable option, believe that you want to *φ*' – that one acquires knowledge regarding one's desires. Byrne provides the following example to motivate his claim:

> The issue of where to dine arises, say. My accommodating companion asks me whether I want to go to the sushi bar across town or the Indian restaurant across the street. In answering that question, I attend to the advantages and drawbacks of the two options: the tastiness of Japanese and Indian food, the cool Zen aesthetic of the sushi bar compared to the busy, garish décor of the Indian restaurant, the bother of getting across town compared to the convenience of walking across the street, and so on. In other words, I weigh the *reasons* for the two options—the 'considerations that count in favor,' as Scanlon puts it (1998:17), of going to either place. These reasons are not facts about my present psychological states; indeed, many of them are not psychological facts at all. (Byrne 2018: 158-159).

We should note that the 'desirable' option here does not mean the 'best' option if 'best' is understood as signifying that option for which we have the most compelling reasons. Byrne is surely correct when he warns that reading 'desirable' to mean 'best' in this sense would lead to DES both under-generating and over-generating cases of self-knowledge (Byrne 2018: 159-160). We can easily know that we desire to eat at the sushi bar, despite not judging it to be the 'best' option in this sense. Further, it is quite possible that though we do not actually desire to go to the sushi bar (perhaps our companion is unpleasant to be around, and we dislike the taste of sushi), we nonetheless judge that going to the sushi bar is best for some practical reasons. This judgement does not move us to follow DES and conclude that we do in fact desire to go to the sushi restaurant.

Rather, 'desirable' is taken by Byrne to mean that which is pleasant, delectable, goodly, or enjoyable. This move is intended to avoid some of the counter-examples just noted. Byrne claims if desirable is understood in the sense of that which is pleasant, delectable, etc., then, for the most part, "one's desires tend to line up with one's knowledge of the desirability of the options; that is, known desirable options tend to be desired" (Byrne 2018: 161). Unfortunately, this amendment does not save the account. Byrne is still forced to acknowledge that DES is defeasible. It remains possible that we might know that something is desirable in the sense specified by Byrne and yet not desire that option. Considering a case of accidie, he writes,

> I know that cycling is desirable yet fail to want to go cycling, but I do not follow DES and falsely believe that I want to go cycling. Lying on the sofa, it is perfectly clear to me that I don't want to go cycling (Byrne 2018: 161-162).

This is problematic for several reasons. As Byrne himself notes, this is a case wherein it seems that we cannot know what we desire by following DES. For, if we were to follow DES, we would end up falsely ascribing the desire to go cycling. Rather, Byrne argues that in cases such as these we recognize that we intend to ψ, that ψing is incompatible with φing, and that ψing is "neither desirable nor all-things-considered better than φing" (Byrne 2018: 165). Even though vegging out on the couch is neither desirable, nor better than cycling, we intend to stay put on the couch and doing so is incompatible with going cycling. Now, this explanation is not particularly helpful since it simply pushes back the problem of accidie. It is possible that we might not know our intentions for the simple reason that we have none. In this case, it would be completely opaque as to why we don't follow DES and falsely ascribe a desire. What is more, DES now seems to rely on whether Byrne's transparency account can provide an explanation of how we come to know our intentions.

Perhaps unsurprisingly, Byrne thinks that it is by means of INT - 'If you will *φ*, believe that you intend to *φ*' – that we can come to know our intentions. But, the now familiar formulation, when applied to intentions, brings formidable problems. Following INT requires endorsing a premise regarding what we will do and perhaps more clearly here than in the case of desire, this would rest on a kind of evidence that is not amenable to a transparency account. In forming a belief about what we will do, we are making a judgment regarding future events. The only evidence for how we will act that we could glean from turning to the world rather than our mental states would be our present or past behavior. From this, we might infer the likelihood of certain actions; however, this method will not be peculiar, since this evidence is available to others in exactly the same way as it is available to us. This is odd given that if any of our self-knowledge is peculiar, it seems to be the self-knowledge we have regarding our own intentions.

Furthermore, INT generates false beliefs in all cases in which there are foreseen, but unintended consequences (i.e., when one recognizes that one will *φ* without intending to *φ*). For example, we may recognize that we will die, without intending to die. In this case, the antecedent of INT will be fulfilled, but it would be false to self-ascribe the intention. To address this problem, Byrne appeals to Anscombe's idea that self-knowledge of one's intentions is arrived at 'without observation'. Byrne takes 'knowledge without observation' to mean 'knowledge not resting on evidence' and suggests that we will not follow INT if we believe that our belief that we will *φ* rests on good evidence that we will *φ* (Byrne 2018: 171). So, though we recognize that we will die, since this recognition is based on substantial evidence, we will not follow INT and conclude that we intend to die. But, as Crispin Wright et al correctly observe, this means that in order to decide whether to follow INT, "one must know on what basis one holds the belief that one will *φ* – a kind of self-knowledge of which

Byrne gives us no account…”.[11] Byrne’s suspect account of how we come to know our desires rests on an equally suspect account of how we come to know our intentions.

5. Problems with Extension

We have so far offered a few reasons to think that two specific iterations of an epistemic transparency account fail to offer an explanation for how we come to know our own beliefs. We have also argued that these accounts are unable to offer an explanation for how we come to know our desires. Indeed, we have seen that Boghossian’s trilemma still poses a challenge for these theories of self-knowledge. Byrne’s way of exploring the inferentialist option has not panned out. Moreover, the prospects of identifying an alternate path are not promising. While Fernández does explore an alternative, he fails to secure a solution. His attempt to satisfy certain intuitions we have about self-knowledge generally, i.e., the intuition that self-knowledge is a cognitive achievement and the intuition that it is privileged, peculiar, and yet fallible, pull in opposite directions. The view becomes incoherent and Boghossian’s trilemma remains intractable.

That being said, our analysis has made one thing clear: part of what motivates Boghossian’s trilemma is a tension between a set of specific intuitions regarding self-knowledge. Given this, it might be worth revisiting some of these intuitions. Must self-knowledge always prove to be privileged and peculiar? Must it always constitute a cognitive achievement? Must our theory of self-knowledge regarding belief extend to other mental states as well? While there are cases to motivate these intuitions, perhaps these cases do not relate to a single overarching type. We want to suggest that self-knowledge is not as uniform as we tend to think and that it is at least possible that these intuitions

[11] P. Conlan, G. Merlo, and C. Wright (2020) “Eyes directed outward. Review essay of Alex Byrne *Transparency and Self-Knowledge*”, *Journal of Philosophy* 117 (6), pp. 332-351.

pertain to distinct types of self-knowledge. If this were right, we might gain a bit of leeway in navigating Boghossian's trilemma. We want to conclude by offering a few reasons to think that a pluralist account is, in fact, viable; or, at the very least, question the feasibility of uniformity.

Though the assumption subtends many approaches to self-knowledge, arguments in favor of uniformity are relatively thin on the ground. Fernández, for his part, stops just short of providing an explicit argument for uniformity, though he clearly hopes to leave open the possibility. He is deterred primarily by differences between propositional and non-propositional states. He writes, "the kind of knowledge that one has of the phenomenal properties of, for instance, one's own sensations and perceptual experience seems to be different from the type of knowledge that one has of the contents of one's beliefs and desires" (Fernández 2013: xii). He goes on to note that self-knowledge of non-propositional states (e.g., pain) is immune from error in a way that our self-knowledge pertaining to propositional states is not. He concludes that providing a theory of self-knowledge with regard to non-propositional states thus constitutes a fundamentally different project.

Unlike Fernández, Byrne does provide an explicit argument for uniformity. The argument consists in two distinct claims. First, Byrne notes that if the epistemology of the mental is not uniform, then certain very specific dissociations would be observed:

> Suppose that a transparency account is correct for knowledge of our beliefs, and that an inner-sense account is correct for desires. Then one would expect to find a condition in which this faculty of inner sense is disabled, sparing the subject's transparent capacity to find out what she believes. Her knowledge of what she believes is similar to ours, but knowledge of her own desires can only be achieved by 'third-person' means (Byrne 2018: 157).

However, Byrne goes on to note that, "no such conditions seem to occur" (ibid.). It is, he claims, reasonable to presume that a single model must account for first-personal knowledge of even the most diverse set of states.

Byrne is surely right that there are no obvious cases in which a subject lacks first-personal access to all of her desires, while access to her beliefs remains intact. However, this does not provide conclusive evidence to suggest that our theory of self-knowledge will be uniform. Indeed, the standard objection to this argument is simply that such dissociations would be difficult to identify since it is often the case that one can gain knowledge of one's desires through third-personal means and this method of gaining self-knowledge might be unaffected by the failure of whatever system allows for first-personal knowledge of our desires.

Byrne's second claim to uniformity is less straightforward. Byrne notes that there are no cases of individuals who have only third-personal access to their mental lives (i.e., lack the self-knowledge in which he is interested) while still retaining rational and epistemic capacities. Based on this, he concludes that epistemic capacities are all that are needed for self-knowledge and this shows that the proper account of self-knowledge must be economical. Since Byrne takes it that the transparency account is the only economical theory of self-knowledge that can explain both peculiar and privileged access, he concludes that "this is a reason for taking it to apply across the board" (Byrne 2018: 158). Notice that even if there are no cases of individuals who have only third-personal access to their mental lives while still retaining rational and epistemic capacities, this is not sufficient evidence for uniformity. The contentious premise here is that the transparency account is the *only* economical theory of self-knowledge available. Suppose that the first-personal knowledge that we have of our mental states, or some subset of our mental states, is gained either in accordance with a transparency model or in accordance with an expressivist model.[12] Since neither theory posits exotic capacities, this hybrid theory would be economical, but this would obviously not entail the uniformity of self-knowledge. Indeed, quite the contrary. We remain neutral on whether this is in fact the case but want to highlight that there is more than one interpretation available for the data Byrne provides. In order to secure his conclusion, Byrne must be able to provide evidence against the idea that there is even a

[12] We are not defending such a model. We are only claiming that it would be equally economical. For a sustained defense of such a model, see Bar-On (2004).

single case of self-knowledge that could be explained by a different economical theory. This is an extremely high bar and one that we do not think Byrne has cleared.

In fact, attempting to tie uniformity to economy in the way Byrne does here actually cuts against the plausibility of transparency accounts. When Byrne provides an account of how we know our memories (Byrne 2018: 183-195), he focuses only on the imagistic aspects of episodic memory. Similarly, when Byrne turns to emotions (Byrne 2018: 172-181), he focuses only on disgust, which is distinctly object-oriented (i.e., we are disgusted by *this* particular thing). This feature of disgust allows Byrne to argue that we exclusively attend to features of the world rather than our mental states when self-ascribing our emotions. However, these examples are not obviously representative. Indeed, it is unintuitive to claim that all emotions are object-oriented, and it is simply false that all aspects of episodic memory are imagistic. At this point, Byrne cannot explain how we come to have self-knowledge in these outlying cases by making reference to additional mechanisms or capacities, as this would mean giving up on the notion of economy. Furthermore, he cannot call upon some alternative, economical theory since this would undercut this second argument for uniformity. Needless to say, this puts a great deal of pressure on the transparency account.

Stepping back from the details of either argument, it is worth asking whether self-knowledge is something that lends itself to a uniform theory. We believe that the sheer diversity of the mental states that can be known first-personally and in a privileged and peculiar way suggests otherwise. Arguably, access to our own beliefs, sensations, emotions, memories, imaginings, thoughts, and perceptual states can be privileged and peculiar. Even without committing to a specific theory of mind, it is easy to see that there is a high degree of diversity among these states. Sensations such as pains and tickles are often described in terms of their characteristic phenomenology though certain propositional attitudes are commonly taken to lack any such phenomenal character. Perception, in addition to its typical phenomenology, is most often described in terms of representational content—content that may be conceptual, non-conceptual, object-involving, and so on. Furthermore, we enjoy a range of emotions, the correct description of which is still highly contentious (see Coliva 2016: 38-47). The states to which we have privileged and peculiar access are diverse, not just in number, but in character as well. We want to stress that this claim regarding the variety of mental states to which we seem to have privileged and peculiar access does not, in and of itself, give us reason to think that a theory self-knowledge *couldn't* be uniform; but, it does suggest that uniformity need not be assumed. Moreover, if our theory of self-knowledge is struggling to make sense of how we can know just one types of mental state (i.e., belief), the call for uniformity seems premature at best.[13]

Throughout this essay, we have been dealing with a species of self-knowledge; namely, self-knowledge as defined by privileged and peculiar access. To be sure, defining self-knowledge in this way delivers up a useful category. But, obviously, not all of our self-knowledge is privileged and peculiar. Take, for example our knowledge of our propositional attitudes. There is a distinction to be drawn here between propositional attitudes as dispositions and propositional attitudes as commitments. Dispositional attitudes are not the direct result of conscious deliberation. A disposition is not the sort of thing that is under our control nor something for which we are thought to be rationally responsible. This is in contrast to commitments, which depend on judgements based on the assessment of evidence and are something for which we can be held rationally responsible.[14] This distinction tracks behavioral differences in the way one relates to one's propositional attitudes such that it is not at all clear that the method by which we come to know our dispositions will look anything like the method by means of which we know our commitments. Both of the attitudes types that were highlighted in this essay (i.e., belief and desire), admit of this distinction. Now, Byrne and Fernández

[13] While this observation regarding the variety of mental states to which we have privileged and peculiar access might only suggest that such a move is premature, it is the conjunction of this observation with our other claims (in particular the subsequent argument regarding dispositions and commitments) that leads us to believe that the pursuit of a uniform account of self-knowledge is ill-advised.

[14] A similar distinction can be drawn with regard to desires as well (see Bilgrami 2006 and Coliva 2016).

are not particularly concerned with this distinction since they are only interested in explaining self-knowledge defined in terms of privileged and peculiar access. Since dispositions are not propositional attitudes to which we have this sort of access, their accounts will not address them. But, note that it is incumbent on us to explain self-knowledge as it pertains to both our dispositions and our commitments. Thus, even if a transparency account were able to offer a uniform account of self-knowledge as it pertains to our propositional attitudes as commitments, some alternative account would still be required to explain how it is that we come to know those very same propositional attitudes as dispositions.

Conclusion

We have shown that the two most prominent epistemic transparency accounts fail to provide a satisfying explanation as to how it is we come to know our own beliefs. Neither seems to successfully navigate Boghossian trilemma. They also fail to fully explain the self-knowledge that we have of mental states other than belief. We focused our critique on desire; however, similar concerns arise when a transparency account is deployed to explain the knowledge we have of our perceptions, memories, emotions, etc.[15] Even if these accounts could be made to work for some of our mental states, there is no reason to think that it could rightfully extend to them all. Moreover, we have suggested that part of what generates Boghossian's trilemma may be a set of conflicting intuitions that have emerged precisely out of assumptions regarding the uniformity of self-knowledge. Regardless, it is clear that, as of now, no uniform theory of self-knowledge is available that can adequately explain self-knowledge. Nor is it obvious that we should seek such a theory. The heterogeneity of states to which we enjoy both privileged and peculiar access suggests that uniformity is likely not a virtue in this domain. If this

[15] For an account of the issues transparency accounts face in explaining belief and perception see Coliva and Mark, (2022).

is right, one of the main perceived advantages of transparency accounts—that is, their promise to provide a uniform account of self-knowledge—is preempted.

References

Bar-On, D. (2004), *Speaking My Mind* (Oxford University Press).

Bilgrami, A. (2006), *Self-Knowledge and Resentment* (Harvard University Press).

Boghossian, P. (1989), 'Content and Self-knowledge', in *Philosophical Topics*: 17: 5-26.

Boghossian, P. (1997), 'What the externalist can know a priori', *Proceedings of the Aristotelian Society* 97: 161-75.

Boyle, M. (2009), 'Two kinds of self-knowledge', *Philosophy and Phenomenological Research* 78: 133-63.

Boyle, M. (2011), 'Transparent self-knowledge', *Aristotelian Society Supplementary* 85: 233-41.

Burge, T. (1996), 'Our Entitlement to Self-knowledge', *Proceedings of the Aristotelian Society* 96: 91-116.

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

Coliva, A. (2016), *The Varieties of Self-Knowledge* (Palgrave).

Coliva, A. and Mark, E. (2020) '*Transparency and Self-Knowledge*, by Alex Byrne', *MIND* 130: 1039-1049

Conlan, P., Merlo, G. and Wright, C. (2020), 'Eyes directed outward. Review essay of Alex Byrne *Transparency and Self-Knowledge*', *Journal of Philosophy* 117 (6): 332-351.

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

Fernández, J. (2013), *Transparent Minds* (Oxford University Press).

McKinsey, M. (1991), 'Anti-individualism and privileged access', *Analysis* 51: 9-16.

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

Moran, R. (2003), 'Responses to O'Brien and Shoemaker', *European Journal of Philosophy* 11: 402-19.

Scanlon, T. (1998), *What We Owe to Each Other*, (Harvard University Press).