Self-Knowledge, Transparency, and Rational Agency

Seoyeon Park

ABSTRACT: It has been observed that intentional mental states are 'transparent' in the sense that we can know our own intentional states by considering only the external states of affairs represented by them. But there is no consensus on how we come to acquire beliefs about our own intentional states transparently and why they should be knowledge rather than mere beliefs. I propose that transparency is grounded in the rational agent's capacity for reason-sensitivity that enables us to engage in reason-based cognitions in accordance with rational norms. These reason-based cognitions presuppose that we have implicit understandings of our attitudes toward the mental contents that occur in them. I argue that it is these implicit understandings from which we come to acquire beliefs about our intentional states transparently. I also argue that these beliefs must be epistemically warranted, and hence be knowledge, if it is to be possible for us to rationally control our minds by reason-based cognitions.

1. Introduction

It is widely accepted that knowledge of our own mental states, or self-knowledge, is special because we have first-person authority over those states in two respects. First, we have *peculiar access* to our mental states in that our way of knowing them is fundamentally different from the ways we know the external world or other minds. Second, knowledge of our mental states is *epistemically secure* in that it is less susceptible to errors than other kinds of knowledge.

Many philosophers have recently proposed that the so-called *transparency* of mental states will shed light on the first-person authority of self-knowledge. The term 'transparency' here does not mean luminosity or self-intimation, but rather characterizes the phenomena in which a person grasps her mental states not by observing her mind but by considering only the external states of affairs that are represented by those states. To borrow Gareth Evans' famous example, someone who is asked the question, "Do you believe there will be a third world war?", will not investigate psychological facts about herself but consider relevant international affairs as if she were given another question, "Will there be a third world war?" (Evans 1982, 225). Evans claimed that perceptions are transparent, too, and it is now widely acknowledged that intentional states including desires and intentions are transparent (Byrne 2018; Boyle 2019).

Transparency is a distinctive feature of first-person attributions, and this is why it is

expected to play a central role in clarifying the first-person authority of self-knowledge. Third-person attributions of mental states to other people are usually made by inferences based on their behavioral evidence. For instance, if you see Jones taking an umbrella, you will infer that he believes that it is raining. In contrast, Evans' example shows that this sort of inference is not needed for first-person attributions. You know *you* believe that it is raining typically by looking out the window. I will call the first-person attributions of one's mental states in this transparent way 'transparent self-attributions' and the resulting knowledge 'transparent self-knowledge'.

But transparency does not directly explain the first-person authority of self-knowledge, for the phenomenon of transparency is mystifying itself. Given that your mental state and the external state of affairs it represents are distinct, how could you know the former by considering only the latter? This question is called 'the puzzle of transparency' (Boyle 2011, 226; Byrne 2018, 77). In fact, this puzzle consists of two different problems. One is *the psychological problem of transparency*, or the problem of clarifying the procedure through which we come to acquire beliefs about our intentional states by considering only external states of affairs. The other is *the epistemological problem of transparency*, or the problem of clarifying why this procedure yields knowledge rather than mere beliefs. Note that these two problems correspond to the requests for explaining the two respects of the first-person authority: A satisfying answer to the psychological problem of transparency should explain why we have peculiar access to our own mental states, and a satisfying answer to the epistemological problem of transparency should explain why self-knowledge is epistemically more secure than other kinds of knowledge.

There are two competing approaches to the puzzle of transparency, which I will call 'the inferential account' and 'the rational agency account', respectively. Proponents of the inferential account hold that transparent self-attributions are the results of making a special kind of inferences, while proponents of the rational agency account hold that they are the results of exercising our rational agency over our own intentional states. My aim in this paper is to

defend and develop a version of rational agency account of transparent self-knowledge by providing answers to both the psychological and the epistemological problems of transparency.

The paper is organized as follows. In Section 2, I will examine the most sophisticated version of the inferential account proposed by Alex Byrne and argue that it fails to solve the epistemological problem of transparency. In Section 3, I will reconstruct Richard Moran's rational agency account in terms of the *capacity for reason-sensitivity*, and criticize that his solutions to the psychological and epistemological problems of transparency are ultimately unsatisfying. In Section 4, I will provide my own answer to the psychological problem of transparency by arguing that our capacity for reason-sensitivity enables us to engage in *reason-based cognitions* in accordance with *rational norms* and that such cognitions presuppose that we have *implicit understandings* of our attitudes toward the mental contents which occur in them. I will then outline how we acquire beliefs about our own intentional states transparently from the implicit understandings. In Section 5, I will suggest an answer to the epistemological problem by arguing that beliefs about our intentional states are epistemically warranted, and hence are knowledge, for our implicit understandings of our attitudes must be epistemically warranted if it is to be possible for us to rationally control our minds by reason-based cognitions.

2. Byrne's Inferential Account of Transparency and Its Problems

Byrne argues that the transparent self-attribution is "an inference from world to mind" (Byrne 2011, 203), and proposes the relevant rule of inference to know one's own beliefs, which he calls "BEL", as follows (Byrne 2018, 102):

BEL If *p*, believe that you believe that *p*.

Following BEL is an inference of which the premise that *p* is about the world and the conclusion that the rule-follower believes that *p* is about the rule-follower's mind. We can then take the procedure of self-attributions as that of making inferences following BEL, so the psychological

problem of transparency is answered. What about the epistemological problem? Byrne attempts to explain why BEL results in knowledge, namely true and warranted[1] belief, by arguing that BEL is knowledge-conducive because following BEL results in a *reliable* and *safe* belief (*Ibid.*, 103-8). A belief is reliable if it is resulted from an instance of a reliable belief-formation process that tends to yield true beliefs, and it is safe if it could not easily have been false (i.e., it is true in all closest possible worlds). Regarding reliability, Byrne claims that BEL is "self-verifying" because anyone who follow BEL cannot but believe the premise that *p* and, accordingly, the belief resulted from following BEL, namely the belief that she believes that *p*, cannot but be true. The process of following BEL always results in a true belief, so the resulting belief is reliable. Also, if a belief resulted from following BEL cannot but be true (even if it were not the case that *p*), it trivially holds that it could not easily have been false. The belief is thus safe.[2]

Further, BEL captures the first-person authority of self-knowledge (Byrne 2018, 109-12). A BEL-instance is a causal transition from believing that *p* to believing that the rule-follower believes that *p*, and this transition is possible only when the rule-follower is the same with the very person to whom the belief *p* will be attributed. BEL thus makes it intelligible that one has peculiar access to one's own mental states. In addition, BEL is self-verifying so that the knowledge resulted from following BEL is less susceptible to errors than other kinds of knowledge. Hence BEL explains the epistemic security of self-knowledge.

But Byrne's inferential account has been forcefully criticized by Matthew Boyle and Moran (Boyle 2011 and 2019; Moran 2012). Boyle argues that following BEL is a "mad" inference (Boyle 2011, 230-1). An inference is a *rational* transition between the contents of premises and conclusion, where the rule-follower is cognizant of the terms and the reasonable

[1] Following the widely accepted usage in epistemology, I use the term 'epistemic warrant' to mean whatever conditions that make knowledge distinct from a true belief. I further follow Tyler Burge in classifying epistemic warrant into two types, *justification* and *entitlement*: justification is epistemic warrant based on propositional reasons, whereas entitlement is the one without propositional reasons.

[2] Byrne actually gives a more complex argument for safety, which I will ignore for the sake of brevity.

relation between them so that she takes the premises as *reasons* to draw the conclusion. But *p* cannot be a reason to conclude that the rule-follower believes that *p*, for the fact that *p* itself does not make it likely that she believes that *p*: There are lots of facts in the world about which one does not have any beliefs. For example, it is the fact that the atomic number of zirconium is 40, but hardly anyone has the belief that the atomic number of zirconium is 40.

Moran similarly argues that BEL cannot be a rule of inference, for it lacks a "rational connection" between the contents of premise and conclusion (Moran 2012, 226-9). Moran contrasts BEL with the rule below, which was initially discussed by Byrne (Byrne 2018, 101):

DOORBELL If the doorbell rings, believe that there is someone at the door.

Following DOORBELL normally yields a reliable and safe, and further epistemically warranted, belief. Moran points out that this warrant force is grounded in the rational connection that the premise of DOORBELL-instance provides a truth-related reason for the conclusion. This connection is shown in the following conditional (Moran 2012, 226):

DOORBELL* If the doorbell rings, there is probably someone at the door.

DOORBELL* is true and can ground DOORBELL's warrant force. If someone recognizes this rational connection, she is epistemically entitled to follow DOORBELL. However, the case of BEL is different. Let us see the following conditional corresponding to BEL (Moran 2012, 227):

BEL* If *p*, you probably believe that *p*.

As we have seen, the mere fact that *p* does not make it likely that you believe that *p*. BEL* is thus false and cannot be a ground for epistemic warrant force. If so, BEL is a baseless instruction, and it is unclear why the belief obtained from following BEL becomes knowledge.

Byrne admits that his account "depends on controversial claims […] that knowledge can be obtained by reasoning from inadequate evidence, or from no evidence at all" (Byrne 2018, 208). But neither Boyle nor Moran has undermined Byrne's arguments that BEL ensures reliability and safety, and he insists that the "mad" inference can yield knowledge as long as

reliability and safety are guaranteed (*ibid*., 124). Yet I think his insistence is problematic. If all we need for a knowledge-conducive rule is reliability and safety, it is not clear whether it is a rule of *inference* from the beginning. Byrne himself remarks that following a rule is to make a judgment *because* a certain condition is obtained, in which this 'because' indicates a "*reason-giving* causal connection" (*ibid*., 101; my italics). But as Boyle and Moran have pointed out, BEL lacks such a rational connection. The premise of BEL-instance offers *no* reason at all for the conclusion. Therefore, following BEL cannot be an inference in the usual sense even though it yields reliable and safe beliefs.

Byrne might introduce another conception of inference, a broad one, that verges on a mere causal shift between two different beliefs so that the rule-follower is permitted to 'infer' although there is no rational connection. However, even in the case that Byrne assumes this broad notion of inference, his account will fail again to solve the epistemological problem of transparency. The claim that your following BEL is making an inference now merely means that you undergo some process, whatever it is, that ensures reliability and safety so that you normally come to know you believe that *p* by considering whether *p*. But if this is all Byrne means while being silent about what grounds the reliability and safety of BEL, it seems that his 'inferential' account is no more than a redescription of the phenomenon of transparency, which is already observed by Evans, and provides no help in understanding the phenomenon.

In sum, Byrne's inferential account of transparency faces a dilemma. If we assume the usual notion of inference and take it as a rational transition between contents, BEL cannot be a rule of inference because there is no rational connection between its premise and conclusion. On the other hand, if we grant a broad sense of 'inference', Byrne's inferential account becomes an empty explanation that is no more than a report that the phenomenon of transparency occurs. Either way, Byrne's account fails to solve the puzzle of transparency.

3. Moran' Rational Agency Account of Transparency and Its Problems

Let us now consider Moran's rational agency account of transparency, according to which it is our rational agency that makes the transparent self-attribution possible. I think that the core of Moran's account lies in his idea that rational agents are "sensitive to reasons" (Moran 2003, 403) and that it is this capacity for reason-sensitivity which connects transparency and rational agency. Moran, however, does not make clear the notions of reason and reason-sensitivity as much as it is desirable, so I will reconstruct his rational agency account while clarifying these notions. Let me begin with Evans' well-known remark on the transparent self-attribution:

> I get myself in a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (Evans 1982, 225)

According to Evans, you can attribute a belief *p* to yourself by the *same* procedure to determine whether *p* is true or not, which is a worldly inquiry. But Evans does not say further about what this procedure is, and Moran suggests that it is one's considerations of reasons:

> The [transparency] claim, then, is that a first-person present-tense question about one's belief is answered by reference to (or consideration of) the *same reasons* that would justify an answer to the corresponding question about the world. (Moran 2001, 62; my italics)

That is, the phenomenon of transparency is the one in which the *reasons* for my second-order[3] judgment that I believe that *p* are *identical* to those for my first-order judgment that *p*. It is not clear exactly what Moran means by 'reasons', but I will suppose that reasons for intentional states are *propositions* that represent external states of affairs.[4] I will also suppose that the rational agency account of transparency is applicable to only those intentional states that can be understood as one's *commitments*. An intentional state, or propositional attitude, consists of

[3] I assume that a 'first-order' intentional state is about external state of affairs and a 'second-order' intentional state is about another mental state of the believer.

[4] This paper does not deal with ontology of reasons, so I will be neutral about what propositions are. If one is unhappy with the very idea that reasons are propositions, one can suppose I use the term 'reason' in a way we use it in ordinary language. For instance, if Jones says, "I believe it will rain because there are dark clouds", the proposition that there are dark clouds is his reason for believing that it will rain.

a proposition and an attitude.[5] The attitudes such as believing, desiring, and intending are one's stances toward the propositions in such ways that taking an attitude is making a commitment about the proposition.[6] For instance, your belief that *p* is your commitment that *p* is true, and your intention to do *φ* is your commitment that you will do *φ*.

I think that Moran's rational agency account will be better understood if we distinguish two kinds of reasons, although he does not make this distinction.[7] I will call them 'normative reason' and 'explanatory reason' for intentional states, respectively. A *normative reason* is a (good) reason to hold an intentional state, which gives an *objective* justification that one makes a *correct* commitment while holding the intentional state. For example, a belief *p* is objectively justified by a normative reason that *q* if *q* supports that *p* is likely to be *true* so that one makes a correct commitment with one's belief *p*. An *explanatory reason*, on the other hand, is one's actual reason for which one holds an intentional state. For instance, if Jones says, "I believe Mary will love this café because it serves nice tea", the proposition <This café serves nice tea>[8] is his explanatory reason for believing that Mary will love this café. An explanatory reason may or may not be a normative one, depending on whether it offers an objective justification.

It is crucial for Moran's rational agency account of transparency that normative reasons have *motivational power* for *rational agents*. If someone is a rational agent, she *will* normally

---

[5] There are well-known debates about whether the contents of perceptual experiences are propositional, but I will put them aside in this paper.

[6] There are exceptions. First, some intentional states such as imaginations seem non-committal. Whether non-committal mental states are transparent is an open question, but I believe that I can extend my account to cover these states as long as they are transparent. However, doing so will be a huge project, and I will focus on beliefs, intentions, and sometimes desires in this paper. Second, some instances of the committal mental kinds are not commitment, e.g., subconscious beliefs and desires. The agent can attribute such states to herself only from the third-person point of view. She is indeed *alienated* from them in that the first-person authority of self-knowledge is not effective, so the states are not transparent to her. I will lay aside the cases of alienation and focus on transparent self-knowledge.

[7] Moran does mention the distinction between *justifying reason* and *explanatory reason* that correspond to *normative reason* and *motivating reason* in the action theory (Moran 2001, §4.4). But this distinction is not equivalent to my classification of reasons in that, typically, a normative reason for an action is regarded as a fact while a motivating reason for an action is regarded as a mental state.

[8] I will use angle brackets ('< >') to denote propositions.

hold intentional states when recognizing normative reasons in favor of them and *will* revise her intentional states when recognizing reasons against them. In consequence, her actual reasons why she holds certain intentional states, namely her explanatory reasons, tend to conform to normative reasons for them. As I see it, *the capacity for reason-sensitivity* is this capacity of the rational agent to hold intentional states based on normative reasons. Because of our capacity for reason-sensitivity, we are responsible for *justifying* our intentional states and this is why we are *rational* agents over them (Moran 2001, 113). If someone believes that *p*, for example, we expect that she will be able to reply to the question, "Why do you believe *p*?", by telling her reasons. In contrast, we do not demand justification for passive mental states such as physical sensations. A question, "Why do you feel the pain on your elbow?", is a value-neutral inquiry for what caused her pain. Even if she cannot answer this question, we will not blame her.

The fact that we have to justify our own intentional states show that we normally follow the principle below, which, as we will see later, Moran himself seems to presume:

> **(Principle of Reason-Sensitivity)** One must have only those intentional states for which one has normative reasons.

This is not to say that one's explanatory reasons for having intentional states must always be normative. But it must be the case that, by and large, the Principle of Reason-Sensitivity holds for rational agents. For if someone were seldom subject to this principle so that recognizing normative reasons which objectively justify (or undermine) her intentional states would *not* normally make any difference to her mind, she will no longer be a rational agent. To review and adjust her own mental states rationally is the *essential* ability of a rational agent, and in this sense the Principle of Reason-Sensitivity is a *constitutive* principle for such an agent.[9]

[9] What I have remarked here is parallel to Donald Davidson's proposal that the (so-called) Principle of Charity is a constitutive principle for language interpretation (Davidson 1973, 324). He argues that we can interpret an alien language *L* only if members of the linguistic community have beliefs that are by and large true and coherent. If this condition is not fulfilled, according to Davidson, we cannot suppose that the native speakers use a meaningful language.

Now, recall Moran's transparency claim that one attributes a belief *p* to oneself by reference to the reasons in favor of *p*. The "reasons in favor of *p*", which support that *p* is true, are normative reasons to believe that *p*. Moran in effect argues that a rational agent can make self-attributions by considering normative reasons for the attributed state. We can see this will naturally follow if the Principle of Reason-Sensitivity is a constitutive principle for a rational agent. It will normally be the case that her explanatory reasons for an intentional state are none other than normative reasons for it. Hence for a rational agent, the normative reasons to hold the belief *p* are also her explanatory reasons why she holds the belief *p*, so they give adequate evidence to answer the psychological question "Do I believe *p*?". But they are *not* pieces of behavioral evidence needed for third-person attributions. Instead, as normative reasons, they are about external states of affairs. Thus, a rational agent for whom the Principle of Reason-Sensitivity is a constitutive principle can know her intentional states by looking outward.

We have so far seen how it is possible for a rational agent to make transparent self-attributions. But the puzzle of transparency remains: The Principle of Reason-Sensitivity itself does not give a solution to the psychological problem of transparency, for it says nothing about the procedure of transparent self-attributions. Nor does it give a solution to the epistemological problem of transparency, for it says nothing about why transparent self-attributions yield knowledge.[10] Moran attempts to solve the two problems, but his solutions are not satisfying.

Regarding the psychological problem, Moran emphasized the role of "deliberative considerations" at first (Moran 2001, 63). In deliberation, a rational agent *forms* an intentional state *as a result of* considering normative reasons. However, though we can and do engage in deliberation, it cannot serve as the general model for self-attributions. As Sydney Shoemaker said, we can know ordinary beliefs such as "I believe I am wearing pants" without deliberation

[10] The Principle of Reason-Sensitivity ensures that second-order beliefs based on the normative reasons are reliable and safe. But I do not assume that reliability and safety are sufficient for epistemic warrant.

(Shoemaker 2003, 395-6). I thus agree with Boyle that transparency applies to a wider range than that of deliberated mental states because one can attribute *pre-existing* states to oneself (Boyle 2019, 1016). To be fair to Moran, I should point out that he does *not* insist that one normally engages in an *explicit* process of deliberation to get self-knowledge (Moran 2001, 63). Though his explanation is not so clear, as I take it, his point was that the *possibility* to deliberate indicates a "categorical difference" between transparent and non-transparent states (*ibid.*, 116). Being able to deliberate about certain intentional states presupposes that the rational agent's consideration on normative reasons will make difference to her states and, in other words, that they are her commitment so that she has the capacity for reason-sensitivity over them. Hence, suggesting the notion of deliberation, Moran in effect claims that transparent mental states are those on which one has the capacity for reason-sensitivity. But if this is the case, he ends up failing to illustrate exactly *how* we make self-attributions based on this capacity.

Let us turn to the epistemological problem of transparency. It is not clear what Moran's solution was in Moran 2001, but in his later works he makes the following remark:

> I would have a *right* to assume that my reflection on the reasons in favor of rain provided an answer to the question of what my belief about the rain is, if I could *assume* that *what my belief* [about the rain] *here is was something determined by the conclusion of my reflection on those reasons* [in favor of rain]. (Moran 2003, 405; my italics and underlines)

Note that the underlined assumption, if generalized, is none other than the assumption that we normally follow the Principle of Reason-Sensitivity (See also Moran 2004, 466-7; and 2012, 231). If the "right" in the above paragraph means epistemic warrant, Moran indeed attempts to solve the epistemological problem of transparency by appealing to the fact that rational agents normally follow the Principle of Reason-Sensitivity. But it is far from clear what it means that we "assume" this fact to engage in self-attributions. It cannot be that we make inferences while taking the Principle of Reason-Sensitivity as a premise, for not only is this implausible itself but also is it incompatible with Moran's claim that transparent self-attribution is immediate

(Moran 2001, 133). If it is not an inference, however, it is hard to see what it could be. Hence Moran's answer to the epistemological problem of transparency is not satisfying, either.[11]

4. How Rational Agents Make Transparent Self-Attributions

In this and the next section I will try to provide better solutions to the puzzle of transparency within the framework of the rational agency account. Regarding the psychological problem of transparency, my solution proceeds in three stages: First, I will argue that the rational agent's *capacity for reason-sensitivity* on intentional states enables her to engage in *reason-based cognitions* in accordance with *rational norms*. Second, I will argue that in order to be capable of making such engagements, a rational agent must not only have explicit understandings of the mental contents that occur in reason-based cognitions but also have *implicit understandings* of her *attitudes* towards those contents. Third, I will suggest that a rational agent who has implicit understandings of her own attitudes toward mental contents will be able to make transparent self-attributions as long as she has required concepts.

Let us begin with the first step. In virtue of the capacity for reason-sensitivity, a rational agent's considerations of normative reasons in "weighing evidence, drawing conclusions, responding to criticism" affect her mind (Moran 2003, 403). Hence this capacity enables her to make cognitive activities by which she can form, assess, and revise her commitments. As we have seen, Moran focused on one type of these activities, deliberation. But I will introduce a more general notion of *reason-based cognitions* that refers to any cognitive activities whereby a rational agent *can* review reasons for her intentional states and sometimes (not necessarily) result in formation, assessment, or revision of them. Asserting a belief, reasoning to a belief, or practical reasoning to an intention all belong to the category of reason-based cognitions.

A rational agent ought to follow certain *rational norms* to engage in reason-based

[11] Shoemaker raises the same doubt in the postscript of his paper. See Shoemaker 2003, 401.

cognitions. As our intentional state is a kind of commitment to the content, rationality imposes on us the norms to improve the likelihood to hold correct commitments. Each kind of attitude is associated with a collection of rational norms. For instance, beliefs ought to be true, so we should follow the rational norms that guide us to true beliefs. Our beliefs ought to be mutually non-contradictory (the Consistency Norm) and support the truth of each other (the Coherence Norm). These norms can sometimes be violated, but such violations indicate flaws in one's rationality. A rational agent's reason-based cognitions ought to be structured in accordance with relevant rational norms in order for her to form, assess, and revise her mental states rationally.

Let us now turn to the second step. To be able to make reason-based cognitions in accordance with rational norms, one must have *implicit understandings* of one's attitudes toward various contents that occur in the cognitions. A reason-based cognition is not a random array of contents, but is organized to be able to justify a specific intentional state. Each content that occurs in the cognition plays a role in the justification, and I claim that what role is played by the content is partly determined by what kind of *attitude* the agent takes toward the content.

I will illustrate what I mean with an example. Jones is majoring in mechanical engineering and has an *intention* to enter the MIT PhD program, and he performs the following practical reasoning: "I will enter the MIT PhD program. For this purpose, I need to make my résumé attractive. Therefore, I will submit one of my papers to the journal *Structural and Multidisciplinary Optimization*." This is an instance of reason-based cognitions in which Jones justifies his intention to submit his paper to the journal. Accordingly, it is organized in compliance with the rational norms related to *intention*. Thus, the proposition <I will enter the MIT PhD program> sets Jones' *goal* for which he decides *what to do*. But the same proposition will play another role if his attitude is different. Suppose, for example, that Jones *believes* that he will enter the MIT PhD program and reasons to justify another belief as follows: "I will enter the MIT PhD program. MIT and Harvard University have promoted lots of academic

exchanges. Therefore, my PhD program will offer me an opportunity to meet Harvard graduate students." In this reasoning, the same proposition <I will enter the MIT PhD program> is not Jones' goal for which he decides what to do but *evidence* to support that the conclusion is *true*.[12]

To engage in reason-based cognitions successfully, one needs to understand what kinds of attitudes one takes toward the contents which occur in them. Imagine that Jones has an intention to enter the MIT PhD program but does not understand that his attitude is that of intending (rather than believing, imagining, etc.) Then he will not particularly care about the rational norms governing intention and thus construct reason-based cognitions where <I will enter the MIT PhD program> occurs in arbitrary ways. He might take this content, which he is intending, as a premise to justify his belief <My PhD program will offer me an opportunity to meet Harvard graduate students>. But this is surely irrational, for no content of intention can justify one's belief or commitment to truth. In this case, therefore, Jones indeed cannot engage in reason-based cognitions in which the content <I will enter the MIT PhD program> occurs.

I thus suggest that a rational agent's reason-based cognitions presuppose not only that she has explicit understandings of the mental contents which occur in them but also that she has *implicit* understandings of her *attitudes* toward the contents. She explicitly understands the contents in that she can recall them and present them in her reason-based cognitions. In contrast, typical reason-based cognitions consist only of propositions that does not make any references to the agent's attitudes, and thus her understandings of the attitudes are not explicit. Still, the rational agent understands her attitudes implicitly because they are reserved in the *structure* of her mind so that they are *implicitly* represented in the structure. For instance, suppose Mary believes that Jones is going to school at 9 a.m. because he said he would. She explicitly

[12] Note that the proposition in question is not the conclusion of both reasonings. The account of Moran might cause a misunderstanding that in order for a mental state to be attributed, the state must be justified in a deliberation so that its content must be the conclusion of a deliberation. But I think that reason-based cognitions presuppose that the agent is not alienated, and thus can know, all the mental states of which the contents correspond to the premises, not only to the conclusion.

understands the two contents, <Jones is going to school at 9 a.m.> and <Jones said he would go to school at 9 a.m.>. These contents are not floating separately in Mary's mind but linked in a way that the truth of the latter coherently supports the truth of the former. This truth-related justifying structure reflects the rational norms that regulate beliefs, so we can see that Mary *believes* both contents although the contents themselves do not contain any elements referring to the attitude of believing. Hence in a rational agent's mind, her explicit understandings of the contents, which can occur in her reason-based cognitions, are organized according to a specific structure that follows relevant rational norms and thereby implicitly represent her attitudes toward those contents. In this sense, she implicitly understands those attitudes, and thus she can organize reason-based cognitions in ways to display this structure.

The final step is to sketch the procedure of transparent self-attributions. I propose that this procedure consists of two stages. First, a rational agent *specifies* the content in her mind. This may or may not be done by deliberation, as we will see.[13] Second, she needs to identify the attitudes she takes toward the content she specifies. But a rational agent who has implicit understandings of the attitudes has already identified the attitude, and what she needs to do is only to *articulate* these understandings. The articulation requires her to possess the first-person concept 'I' and the relevant psychological concepts so as to apply a complex concept such as 'I believe' to the specified content. Without these concepts, one cannot get self-knowledge. For instance, an autonomous car may be credited with an intention to go along the optimal route and accordingly with an implicit understanding of its attitude of intending. Still, the car will be unable to judge, "I intend to get on the highway", unless it has the concepts 'I' and 'intend'.[14]

[13] Moran saw deliberation as the complete procedure of self-attribution, so his account clarifies the stage of content specification but misses that of attitude identification. My notion of implicit understanding fills up the gap and makes it intelligible how one comes to know both the content and the attitude.

[14] I am inclined to be skeptical that an autonomous car or an AI robot at present can have intentional states in the genuine sense beyond displaying stimulus-respond patterns. Yet my point is that it *would* have implicit understandings of its attitudes toward the contents of its intentional states *if* it has them.

Let us see an example of the *deliberative* transparent self-attribution of a belief, where one specifies the content of the belief by deliberation. Suppose Jones asks himself, "Do I believe that Mary will go to school on Friday?". To determine whether Mary will go to school on Friday, he would consider external states of affairs and perform a reasoning as follows:

(P1) There is the Solid Mechanics final exam on Friday.

(C1) Therefore, those who take the course of Solid Mechanics will go to school on Friday.

(P2) Mary is taking the course of Solid Mechanics.

(C) Therefore, Mary will go to school on Friday.

This is a reason-based cognition where (P1), (C1), and (P2) together give a normative reason that objectively justifies the belief whose content is the conclusion, (C). As Jones is a rational agent for whom the Principle of Reason-Sensitivity is a constitutive principle, he will come to believe (C) based on this reason and thus the above reasoning is his deliberation. Jones judges that Mary will go to school on Friday as a result of the deliberation, so he specifies the content <Mary will go to school on Friday>. Once this content is specified, he needs to identify his attitude toward it as that of believing (rather than wanting, intending, etc.). This can be done much easily. The above reasoning was possible because Jones had implicit understandings of his attitudes toward the contents that occur in it. Thus, he is indeed ready to grasp what attitude he takes toward <Mary will go to school on Friday> once he specifies this content. If Jones has the concepts of 'I' and 'believe', he can articulate his implicit understanding of the attitude toward the content and hence immediately comes to know that *he believes* that Mary will go to school on Friday. This self-attribution is transparent because in the above reasoning, Jones only attends to the contents which represent external states of affairs, not his mental states.

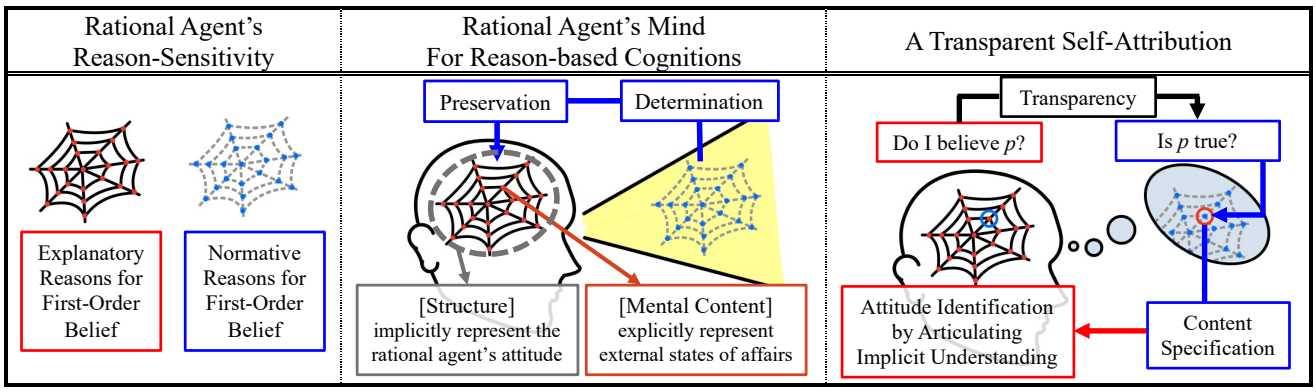
Yet performing a deliberation is not necessary for a transparent self-attribution. Jones can make a *non-deliberative* self-attribution in which he specifies the content directly, i.e., without actually considering normative reasons, and then identifies his attitude. The difference

between deliberative and non-deliberative self-attributions lies in the stage of specifying the content. In the former case, the content is not yet fixed for Jones, so he needs to deliberate for specifying it. By contrast, in the latter case, the content is directly available to Jones so that he can be aware of it without any explicit reason-based cognition. But it is important that the procedures of two cases are not fundamentally different. In both cases, the mental state that Jones will attribute to himself *can* occur in his reason-based cognitions, which implies that he has the capacity for reason-sensitivity over the state. It is this capacity that enables him to specify the mental content without observing his own mind and also that offers him the implicit understanding of his attitude toward the content, in virtue of which he immediately identifies the attitude regardless of whether the content is specified deliberatively or non-deliberatively.

The suggested procedure of transparent self-attribution makes good sense of the first aspect of the first-person authority of self-knowledge, namely *peculiar access*. A rational agent has the capacity for reason-sensitivity only over her own mind, so the Principle of Reason-Sensitivity is effective exclusively in the first-person perspective. If you find normative reasons to believe that *p*, they will be your explanatory reasons for which you believe that *p*, but there is no guarantee that they will be other people's explanatory reasons. Hence in the way I have proposed, you cannot specify the content that is specified by other people. In addition, you may have no understanding at all about other people's attitudes toward the content. Therefore, the transparent self-attribution is possible only from the first-person point of view.[15]

[15] Boyle has recently suggested an account of transparency that has interesting points of contact with mine (Boyle 2019). He claims that our awareness of the world in our intentional states always involves our *implicit awareness of the manners* in which we are aware of the world. Further, he argues that each manner in which the subject is aware of the world corresponds to a specific kind of mental state she is in, and that the subject can make the manner explicit by self-ascribing the state through *reflection*. Boyle's claim that one has implicit awareness of the manner in which one is aware of the world seems similar to mine that a rational agent has implicit understandings of her attitudes toward contents. But he characterizes this implicit awareness in terms of Sartre's notion of "non-positional consciousness", which is supposedly a consciousness of the very consciousness itself and does not posit any object it is about (*ibid.*, 1029). I find this notion mystifying: how could consciousness be consciousness of itself without positing itself as an object? Also, how could reflection, whatever it is, transform such "non-positional consciousness" into explicit self-knowledge?

The following pictures summarize the main concepts suggested in the Section 3 and 4.



5. Why Transparent Self-Attributions by Rational Agents Yield Knowledge

Let us now address the epistemological problem of transparency. My answer to this problem within the framework of the rational agency account is heavily indebted to Tyler Burge's idea that self-knowledge obtains its epistemic status due to the *rationality* of relevant cognitive activities in adjusting one's mind (Burge 1996). Before exploring his argument, I want to emphasize that the main issue here is *epistemic warrant* while knowledge is true and warranted belief. For it is clear that transparent self-attributions, illustrated in Section 4, normally yield true beliefs. An attributed mental state consists of a proposition the agent explicitly understands and an attitude she implicitly understands. The agent must hold the mental state to possess the understandings, so her second-order judgment cannot but be *true*. In addition, if transparent self-attribution is successfully done, the subject cannot but *believe* the resulting judgment.

Burge introduces the notion of *critical reasoning*, namely the reasoning in which the reasoner not only makes ordinary reasoning but, at the same time, evaluates whether her reasons are reasonable (Burge 1996, 73-4). He thinks that rational agents can normally perform critical reasonings. Here is an example:

(P1) I believe that the school cafeteria is concerned about health.

(P2) I have gained a reason to believe that the school cafeteria serves leftovers again.

(C1) Therefore, I believe that the school cafeteria serves leftovers again.

(P3) That the school cafeteria serves leftovers again is a reason against believing that the school cafeteria is concerned about health.

(P4) There is no sufficient reason that counteract this reason against believing that the school cafeteria is concerned about health.

(C) Therefore, the school cafeteria is not concerned about health.

A rational agent can form, assess, and revise an intentional state by critical reasoning. In this respect, critical reasoning is similar to my notion of reason-based cognition. But critical reasoning is significantly different from the reason-based cognition in that it not only involves first-order propositions that represent external states of affairs but also second-order ones that represent the reasoner's intentional states as its premises.

Burge argues that our epistemic warrant to self-knowledge is grounded in the fact that self-knowledge is integral to critical reasoning (Burge 1996, 75). His argument can be taken as a *reductio ad absurdum*: Suppose that a rational agent's beliefs about her intentional states are *not* epistemically warranted. This means that her second-order beliefs would largely fail to represent her intentional states in a reliable way. If such an agent engages in critical reasoning whose premises refer to her first-order intentional states, they will not be reliable references to review those first-order states. Hence the resulting formation, evaluation, or revision of her first-order intentional states would become arbitrary and irrational. If so, however, she cannot be a rational agent, which is contradictory to our original supposition. Therefore, the rational agent's ability to perform critical reasoning transcendentally requires that her second-order beliefs about her first-order mental states are epistemically warranted. This warrant is an (*a priori*) entitlement, not justification, for it relies on no specific propositional reason but only on the general (*a priori*) assumption that our critical reasoning is by and large rational.

The above argument of Burge implies that we need to possess the ability to engage in critical reasoning in order to be able to acquire self-knowledge. It seems to me, however, that

the former ability requires more than the latter. I thus find that Burge's argument is problematic because it makes his account of self-knowledge a *hyper-intellectualized* one.[16] For example, suppose that Mary finds that the school cafeteria is serving the same menu as yesterday. She attributes a belief to herself and obtains self-knowledge, saying "I believe that the school cafeteria offers me leftovers". Yet she may be unable to see that this belief provides a reason to undermine another belief of her that the school cafeteria is concerned about health. Unless she comprehends her self-knowledge as a reference to assess and modify her first-order intentional state, she could not perform critical reasoning. But this requires a more complex ability than the one to hold self-knowledge.

Burge have emphasized critical reasoning probably because he believes that we must be able to perform critical reasoning to control and adjust our intentional states rationally. I think, however, we can do so by engaging in reason-based cognitions that do *not* require self-knowledge. That is, even though someone lacks concepts such as 'belief', 'is a reason for', etc., and thus cannot obtain self-knowledge, she can still weigh reasons and modify her intentional states in reason-based cognitions. For example, a small kid can undergo the following line of thought: "It's unlikely that I will go out for dinner tonight. But wait! Today is my brother's birthday. And my family has always eaten out on the birthday of a family member. So I will go out for dinner tonight." It is simple reasoning where her initial belief <I will probably not go out for dinner tonight> is revised into a well-justified belief <I will go out for dinner tonight>. She is assessing and modifying her belief rationally, although she may not possess the concept of belief and thus may not be able to acquire self-knowledge.

Still, Burge's *reductio* argument contains a crucial insight that the epistemic warrant for self-knowledge is grounded in the fact that we have an ability to adjust our intentional states

[16] This is ironical, given that Burge is well-known for criticizing a number of philosophers that they have had the tendency of hyper-intellectualizing our capacities for perceiving, believing, speaking language, etc.

rationally. In his picture, it is critical reasoning that manifests this ability, to which our self-knowledge is an integral part. I claim that the significance of implicit understandings in reason-based cognitions is parallel to that of self-knowledge in critical reasoning. In the above example, the kid might lack the psychological concept of 'belief', but her reasoning was possible because she *implicitly understood* that her attitudes toward the contents <I will probably not go out for dinner tonight> and <I will go out for dinner tonight> were those of *believing* (rather than intending, imagining, etc.). I think that the epistemic warrant for such implicit understandings is grounded in the fact that they are an integral part of reason-based cognitions and that we have an ability to adjust our intentional states rationally in those cognitions. The warrant will be transmitted to self-knowledge after we successfully make transparent self-attributions.[17]

In sum, the capacity for reason-sensitivity transcendentally requires a rational agent to be epistemically entitled to her implicit understandings. She can form, assess, and revise her mental states by engaging in reason-based cognitions. This is because the Principle of Reason-Sensitivity is a constitutive principle for her, so it must be the case that her considerations about (normative) reasons will actually affect her mind. The ability to engage in the cognitions presupposes that she implicitly understands her attitudes toward the contents that occur in them. Now, similar to Burge's argument, if it is to be possible for reason-based cognitions to function well in adjusting her minds rationally, the implicit understandings must be epistemically warranted. Otherwise, the cognitions will result in irrational assessments and modifications.

The above argument for our epistemic warrant to self-knowledge explains why self-knowledge is epistemically more secure than other kinds of knowledge, for this warrant ensures what Burge calls "the immunity to brute error" (Burge 1996, 81-3). A *brute error* is the one

[17] One may wonder how implicit understandings, which are non-propositional and hence cannot be true or false, could be epistemically warranted. Let us stipulate that such understandings are *veridical* when the structure of our mental states properly reflects the kinds of attitudes taken by the agent. We can then extend the notion of epistemic warrant as follows: An implicit understanding is *epistemically warranted* if it comes from a good route for veridical understanding.

that does *not* derive from short of rationality or cognitive malfunction. If brute errors are permitted, it can be the case that a belief is not true but still epistemically warranted. Our sensory organs permit brute errors in perceptual beliefs: For example, a traveler may falsely believe that there is an oasis on the right side, seeing a mirage in a desert. The belief is false, but she is still epistemically entitled to the belief because she looks at an illusion which is visually indiscernible from a real oasis. By contrast, our implicit understandings of our own attitudes cannot but be veridical if they are epistemically warranted. For warranted implicit understandings are integral to reason-based cognitions and if the implicit understandings are not veridical, the cognitions will be irrational. Thus, our implicit understandings, and our self-knowledge obtained from them, are immune to brute errors. As knowledge about the external world and other minds is not, transparent self-knowledge is epistemically more secure.

6. Conclusion

I have argued that we can form and be entitled to transparent self-knowledge because of our rational agency in the sense of reason-sensitivity. A rational agent has the capacity for reason-sensitivity that leads her to hold intentional states on the basis of normative reasons. This capacity enables her to engage in reason-based cognitions which presuppose that she implicitly understands her attitudes toward mental contents that occur in them. With the capacity for reason-sensitivity, she can specify a content by considering normative reasons that represent external states of affairs. She can then identify which attitude she takes toward the content in virtue of her implicit understanding of the attitude. The rational agent must be epistemically entitled to the implicit understandings of her attitudes, and thus to the self-knowledge acquired from them. If not, her reason-based cognitions will result in irrational assessments and revisions of her mind in a large scale, which is impossible given the fact that she is a rational agent.

Reference

Alvarez, Maria. (2017). "Reasons for Action: Justification, Motivation, Explanation", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.

Anscombe, G.E.M. (2000). *Intention*. 2nd Edition. Harvard University Press.

Boyle, Matthew. (2009). "Two Kinds of Self-Knowledge". *Philosophy and Phenomenological Research* 78 (1):133-164.

____. (2011). "Transparent Self-Knowledge". *Proceedings of the Aristotelian Society Supplementary Volume* 85 (1):223-241.

____. (2019). "Transparency and reflection". *Canadian Journal of Philosophy* 49 (7):1012-1039.

Burge, Tyler. (1996). "Our entitlement to self-knowledge". Reprinted in Burge, Tyler (2013). *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3*. Oxford University Press UK: 68-87.

____. (1998). "Memory and Self-Knowledge". Reprinted in Burge, Tyler (2013). *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3*. Oxford University Press UK: 88-103.

____. (2011). "Lecture II: Self and Constitutive Norms". Reprinted in Burge, Tyler (2013). *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3*. Oxford University Press UK: 166-186.

____. (2011). "Lecture III: Self-Understanding". Reprinted in Burge, Tyler (2013). *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3*. Oxford University Press UK: 187-226.

____. (2013). "Introduction", *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays, Volume 3*. Oxford University Press UK: 1-52.

Byrne, Alex. (2011). Transparency, belief, intention. *Aristotelian Society Supplementary Volume* 85:201-21.

____. (2018). *Transparency and Self-Knowledge*. Oxford University Press.

Davidson, Donald (1973). "Radical interpretation". *Dialectica* 27 (1):314-328.

Evans, Gareth. (1982). "Self-Identification". *The Varieties of Reference*. Oxford University Press: 205-258.

Gertler, Brie (2011). "Self-Knowledge and the Transparency of Belief". In Anthony Hatzimoysis (ed.), *Self-Knowledge*. Oxford University Press: 125-145.

____. (2020). "Self-Knowledge". *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). Edward N.Zalta (ed.). URL = https://plato.stanford.edu/archives/spr2020/entries/self-knowledge/

Keeling, Sophie (2019). "Knowing our Reasons: Distinctive Self-Knowledge of Why We Hold Our Attitudes and Perform Actions". *Philosophy and Phenomenological Research* (2):318-341.

Moran, Richard. (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press.

____. (2003). "Responses to O'Brien and Shoemaker". *European Journal of Philosophy* 11 (3):402-419.

____. (2004). "Replies to Heal, Reginster, Wilson, and Lear". *Philosophy and Phenomenological Research* 69 (2):455-472.

____. (2012). "Self-Knowledge, 'Transparency', and the Forms of Activity". In Declan Smithies & Daniel Stoljar (eds.), *Introspection and Consciousness*. Oxford University Press. 211-236.

O'Brien, Lucy. (2003). "Moran on agency and self-knowledge". *European Journal of Philosophy* 11 (3):375-390.

Shoemaker, Sydney. (2003). "Moran on Self-Knowledge". *European Journal of Philosophy* 11 (3):391-401.