Self-consciousness — In a Naturalistic Framework

Roberto Horácio Sá Pereira[1]

1 Universidade Federal do Rio de Janeiro

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

In this essay, I present a brand-new naturalistic model of self-consciousness that draws inspiration from Dretske's program for naturalizing the notion of representation. Dretske's naturalization program is very ambitious in many respects. However, one of the crucial aims is overlooked in the literature, namely the abandonment of the traditional conception of representational content as a "Vorstellung," in which the individual occupies both poles of representation: the represented subject and the represented object. Just as sensory states containing information about a mind-independent source are "recruited" by selection with the function of representing that source without a subject, an "EGO file" housing autobiographic information is "recruited" by selection with the function of representing the cognitive system itself without a subject. The most basic form of self-reference is not the result of an act carried over by an "underlying subject:" The subject does not need to occupy both poles of the representation, the subject and the object poles. Instead, it is the result of the "recruitment" of the autobiographic memory with the function of representing the subject itself.

Keywords: Self-consciousness; Knowing Self-reference; The Puzzle of Self-consciousness.

> ***"Wo ES war, soll ICH werden"***
> ***(Sigmund Freund)***

Prolegomena

"Self-consciousness" is a technical term introduced in modern philosophy (probably in the Leibniz-Wolf-Kant tradition through the word "apperception"). There is no equivalent in the ordinary colloquial usage in the Indo-European languages that captures the meaning of the word as employed in philosophy. In the colloquial speech, the word "self-consciousness" has a much more restricted meaning. In the English language, the colloquial meaning of "self-consciousness" is a heightened awareness of one's actions, appearance, or behavior and an accompanying sense of being hindered and embarrassed.

Interestingly, "self-conscious" in German (“selbstbewusst") has a positive sense of self-worth, meaning essentially self-confidence. In the Romance languages, "self-conscious" is colloquially understood as a person who acts reflectively, thinking about the consequences of her actions for both her and others. A self-conscious individual is a person who is aware of their integration into the social environment and acts in accordance with his or her interests or in accordance with moral interests. The individual without self-consciousness is essentially alienated. In all Indo-European languages, "self-consciousness" refers to forms of the relation of oneself to oneself that concern one's behavior in relation to others. None of these colloquial meanings in the Indo-European languages is the focus of this paper.

To Locke, however, some human beings are conscious not only of themselves concerning their behaviors in relation to other human beings. They are conscious of the world that surrounds them but also conscious of themselves as individuals: their activities, their bodies, and their own mental states. The consciousness they have of themselves in the cognitive sense is what philosophers call self-consciousness in the technical sense:

> *We have an intuitive Knowledge of our own Existence, and an internal infallible Perception that we are. In every Act of Sensation, Reasoning, or Thinking, we are conscious to ourselves of our own Being. (1700: IV.ix.3)*

Locke postulates that self-consciousness is a key element in determining the concept of an individual, whereby a person is an "intelligent being possessing reason and introspection and able to know oneself as the same thinking entity at different times and places" (1700: II. xxvii.9). Self-consciousness is also an important factor in his concept of personal identity (see §4.1). Whether self-consciousness was the subject of philosophical debate in ancient and medieval philosophy is an open question that does not concern me here. One thing is certain: self-consciousness has been a central concept in modern philosophy since Descartes. Indeed, self-consciousness has been a concept in modern philosophy since the Cartesian writings of the early 17th century. It is the ability to recognize oneself as an individual separate from others, as well as the awareness of one's own thoughts, feelings, and actions. Since Descartes, philosophers have considered self-consciousness to be the foundation of human knowledge and the basis for our ability to think and make moral decisions.

However, if self-conscious beings are aware of themselves, their activities, bodies, and mental states, why does self-consciousness in the philosophical, technical sense of being immediately conscious of oneself become an enigma? The key problem of self-conscious in the modern philosophy of mind being is self-conscious is not self-reference by accident. Self-consciousness involves *knowing* that one is referring to oneself. Consider the pair of thoughts:

1. Oedipus thinks that Laius's feels ashamed.
2. Oedipus thinks that he* feels ashamed.[1]

In (1), Oedipus is certainly referring to himself, albeit unwittingly—he is not aware of it. In (2), by contrast, Oedipus is referring to himself with full knowledge. Therefore, being accidentally aware of oneself is different from being self-conscious in the sense of intentionally referring to oneself. It is only the latter, the so-called "de se" sense, that is pertinent for the philosophical debate. The fundamental problem is how to account for the difference between (1) and (2), avoiding

an infinite regression or a vicious circle.

My main source of inspiration is Dretske's program for naturalizing the mind, and the key concept is that of representational content. My point is that one of the crucial achievements of the naturalization of representation is overlooked in the vast literature, namely the abandonment of the traditional conception of representational content as a "Vorstellung," in which the individual occupies both poles of representation: *the representing subject and the represented object.* Just as sensory states containing information about a mind-independent source are "recruited" by selection with the function of representing that source, an "EGO file" housing autobiographic information is "recruited" by selection with the function of representing the cognitive system itself without a subject.

My argument is the inference to the best explanation. If there is a solution to Fichte's puzzle, the only solution is mine. The most basic form of self-reference is not the result of an act carried over by an "underlying subject:" The subject does not need to occupy both poles of the representation, the subject and the object poles. Instead, it is the result of the "recruitment" of the autobiographic memory with the function of representing the subject itself.

I have structured this article as follows: After this introduction, the next section (Section 1) is devoted to the formulation of Fichte's old riddle. The next two sections (Section 2 and Section 3) are dedicated to the introduction and evaluation of Tugendhat's famous "linguistic-analytic" approach to the puzzle of self-consciousness. I have chosen Tugendhat's view (1979) not only because it was a great success in Germany at the time but also because it was intended as an alternative to the traditional approach based on the concept of representation as "Vorstellung" and the so-called theory of reflection.

In the fourth section, I will assess Bermúdez's and the traditional phenomenological accounts. Bermúdez's modest claim is to solve what he calls "the paradox of self-consciousness," namely the explanation of the mastering of the token-reflexive role of the employment of the first-personal pronoun, namely by appealing to the first-person nonconceptual contents. In contrast, the Sartrean phenomenological tradition aims to account for ordinary, full-fledged self-consciousness by appealing to a putative form of pre-reflexive or intransitive self-consciousness. I argue that all proposals are doomed to fail because they assume the concept of representations as "Vorstellung." Finally, in the last section, I propose my solution, namely, the naturalization of self-representation.

I. *The Riddle of Self-consciousness*

The idea that consciousness requires representation is quite plausible. For example, I am aware of this flower in my garden because I can represent it by seeing it. Conversely, I become aware of my visual experience of a flower in my garden because I represent my first-order visual representation (through thought or "inner perception") via a second-order representation (either a higher-order thought or a higher-order perception). In the history of modern philosophy, Locke was certainly the first to use the word "reflection" to denote the capacity of our mind to turn its gaze upon itself and make its processes the object of its consideration (Locke, 1690). Since then, it has been common to refer to such a higher-order account of self-consciousness as the reflection theory of self-consciousness. Moreover, Descartes, Leibniz, Wolf and Kant have endorsed the theory of reflection. Be that as it may, the so-called theory of reflection goes back to Ockham.

However, as we have seen in the Prolegomena, in its technical sense, a self-conscious being is not just a being accidentally aware of itself, as one would be if looking at an old photograph without recognizing its image. Rather, a self-aware being knows that it is referring to itself. Perry illustrates this case with a famous example:

3. The shopper with the torn sack is making a mess.
4. I am making a mess.[2]

However, what about self-consciousness, the consciousness of myself as myself? Just as I become aware of the flower by representing it, of my experience by meta-representing it, I become aware of myself through a second-order representation of myself as the subject and object-pole of my first-order representation. David Rosenthal (2004) is one of the several authors who have reformulated the old theory of reflection into his higher-order theory. According to this theory, self-consciousness arises when the subject of a first-order thought is "disposed to have another [second-order] thought that does so identify the individual the first thought is about" (2004, p. 167). In this way, the *subject* of the first-order representation becomes *the object* of the second-order representation.

We have here the two traditional views that engender the puzzle of self-consciousness. First, representation is traditionally conceived as a "Vorstellung" in the traditional sense, wherein it always comprehends subject and object poles. Second, self-consciousness is conceived as the result of reflection: the subject of a first-order representation is disposed to entertain a higher-level representation that identifies itself as the object of the first-order representation. In this context, the strange Fichtean statement "I = I" ("ich = ich") seems to make sense: the "I," as the subject of the second-order thought, identifies itself as the object of the first-order thought. But this is precisely why the puzzle arises. In order to identify myself as the subject and the object ("ich = ich"), I tacitly presuppose that I know beforehand that I, as the subject of the second-order thought, represent myself* (the subject of the first-order thoughts) as the object by the second-order thought.

Fichte was the first to recognize an enigma when he attempted to clarify self-consciousness through a reflexive act of recognition (Henrich, 1966, p. 14).[3] In terms of the theory of reflection, knowing self-reference presupposes the knowledge that the object of reflection is, at the same time, the subject who performs the act of reflection. Fichte sums up the paradox of self-consciousness by complaining that "In this way, however, our consciousness is not explained (...) one assumes it to be a state of mind or an object and thus always presupposes a subject, but never finds it" (Fichte, 1982, ii, p. 356).[4]

Fichte's conundrum is this: The knowledge of self-reference requires the knowledge of the identity between the subject as thinker ("Is") and the subject as object ("Io"), the thinker as subject and object at the same time, hence the seemingly paradoxical Fichtean sentence: "Ich=Ich." The supposed riddle can be articulated more clearly in a classical quandary. The first branch of the conundrum is the infinite regress. The question is: How am I to know I am the object of my reflection? The answer is by recognizing that I am the one who performed the act of reflection in the first place. However, the same question arises again: how do I know I am the subject who performed the first-order reflection? (How can I be sure that I am the source of the corresponding token of the first-person pronoun?) To do this, I would need to perform another second-order reflection to verify myself as the subject who performed the first-order reflection, and so forth *ad*

infinitum.[5]

Suppose you are suffering from severe pain. You pick up your voice to relieve your suffering and say, "I am in severe pain!" Years pass, and you find the tape on which your voice is heard, "I am suffering from severe pain." However, perhaps because people do not easily recognize their own voices when it is recorded, you do not identify yourself as one who was in severe pain. The point is, were you not aware of yourself when you said, "I am in severe pain"?" in the past. Moreover, when you think of your voice on the tape, are you not self-consciously thinking? What role does higher-order representation play in self-consciousness?

Suppose now I want to avoid the undesirable vicious infinite regress. In this case, to detain the infinite regress, I must assume that I somehow know in advance that I am the same individual who is performing the act of reflection and the object of my own reflection. However, if I already know that I am the same individual who is reflecting the subject and the object of my own reflection, as Fichte complaints: self-consciousness is not explained but presupposed (1974, p. 56.). This is what Fichte calls a vicious circle.

Fichte's solution to this problem is rather opaque: "self-positing."[6] According to Henrich, Fichte never explained his metaphor of positing and self-positing (1966, p. 18). The formula "the 'I' posits itself up" can only negatively characterize Fichte's rejection of the theory of reflection. On the other hand, the idea of "self-positing" sounds incomprehensible to the Heidelberg School. Pothast wonders, "How could someone perform this act of positing if he does not yet exist in the first place" (1971, p. 71)?

Thus, the Heidelberg theory of consciousness emerged in 1971 as an attempt to resume Fichte’s original insight. Negatively, it can be characterized by the rejection of both the old theory of reflection and Fichte’s claim that self-consciousness is a sort of intellectual intuition of the sheer activity of apperception. Positively, it can be seen as the resumption of Fichte’s original insight that self-consciousness must be based on a non-propositional knowledge of oneself, which Henrich calls self-acquaintance.

Three main theses can summarize the core of the old Heidelberg theory of self-consciousness. (1) The old theory of reflection cannot explain knowing self-reference without circularity. (2) To break the circle, self-consciousness must be explained based on an original form of self-acquaintance within consciousness. (3) This original form of self-acquaintance is neither an activity nor a relation between a subject and its object.

The main problem with the old Heidelberg theory is this: Even if one abandons the traditional theory of reflection, i.e., the idea that self-consciousness arises as the result of a meta-representation of oneself as the object of a first-order representation, it still clings to the traditional idea of representation as "Vorstellung," comprehending the subject pole—the representing subject—and the object pole—the represented object. Under the assumption of representation as "Vorstellung," it remains a mystery, however, what exactly (2) and (3) mean and how they are supposed to solve the old riddle. It remains a mystery what self-acquaintance can mean and how the self is self-acquainted with herself.

II. *Language-analytic Approach*

Tugendhat claims that Fichte's conundrum arises because the philosophical tradition misunderstands self-consciousness within the traditional framework of the "Vorstellung," which he calls, following Heidegger, the subject-object model. To illustrate, I became aware of this computer by intentionally representing it as an object. Likewise, I should become aware of myself as the same individual who is the object of a second-order representation and the subject who performs this second-order representation. Tugendhat summarizes his critique of the Fichtean puzzle of self-consciousness as follows: "The problem that Henrich identifies... rests on the assumption that we are analyzing something whose essence consists in the identity of knowing and what is known" (1979, p. 64).[7]

According to Tugendhat, the root of the problem of the old reflection theory is the subject-object framework. The conundrum arises because the tradition misunderstands self-consciousness as the alleged identification relation between the "I" as the representing subject and "I" as a represented object that results from self-representation (Fichtean sentence: "Ich=Ich"). The underlying assumption is that one becomes aware of oneself by identifying oneself as the object of one's intentional act of representation. In this respect, Tugendhat is correct. We will revisit this issue in the final section.

Tugendhat's "linguistic analytic approach" is characterized by two closely related principles. The first is his rejection of the subject-object framework. The second is his "language-analytic" reduction of the phenomenon of reflexive self-reference to the mode of employment of psychological I-sentences in which one attributes to oneself a mental predicate “ϕ.” Consequently, to understand knowing self-reference, one needs to understand the way of using the first-person pronoun and mental predicates.[8]

As the ultimate referent, the first-person pronoun does not identify me or single me apart from others in a particular field. The real question is why. Unlike Wittgenstein and Anscombe, Tugendhat claims that the "I" refers to my person as an individual identifiable from a third-person perspective. As a result, every sentence "I ϕ" expresses a genuine proposition rather than merely an avowal. On this basis, Tugendhat formulates his principle of veridical symmetry: "The sentence "I ϕ" is true, if uttered by me, *iff* the sentence "He ϕ" is true if uttered by someone else who by "he" means "me" (Tugendhat, 1979, p. 88).

According to Tugendhat, what ensures veridical symmetry is the reasonable assumption that the first-person and third-person indexicals involved may co-refer. When someone refers to himself with the first-person pronoun and when someone else (or the person himself) refers to that person with the third-person pronoun:

5. Oedipus feels ashamed.
6. I think feels ashamed.

The simple co-reference of the indexicals involved is sufficient for the veridical symmetry between first and third-person psychological sentences about the same person. Tugendhat overlooks, however, that what guarantees the veridical symmetry is rather a coarse-grained way of individuating the proposition in question, namely, the Russellian view of propositional content in which the proposition that Oedipus feels ashamed is an ordered set of two items whose members

are {Oedipus; the property of feeling ashamed}.

Although 5 and 6 express the same coarse-grained proposition, it is only by thinking 6 that Oedipus knows that he refers to himself without identifying himself. Unlike 5, Oedipus's or any other person's knowledge of the truth of 5 is based either on observation of Oedipus's behavior (if the thinker of 5 is another person) or, in some cases, on inference.

But if the immediate knowledge of oneself as the owner of mental states is negatively described as not based on observations, inferences, and alleged inner perception, Tugendhat owes us a positive explanation of it. Following Wittgenstein and Shoemaker, Tugendhat holds that psychological first-person sentences are immune to a peculiar error of reference when employed in conformance to the rule. So, if Oedipus knows the rule of employment of the first-person pronoun (according to which that pronoun refers to whoever employs a token of it), by employing a token of it, Oedipus couldn't possibly fail to recognize that he is referring to himself whenever he thinks the content of sentence 6.

Yet, Tugendhat's equation of immediate epistemic self-consciousness and the employment of psychological I-sentences in conformity to its rule raises several questions. First, what guarantees the immediate self-knowledge of the content expressed by 6 is certainly not the Russellian proposition consisting of the sequence {Oedipus; feels ashamed}, but rather the mastering of the token-reflexive rule of the employment of the first-personal pronoun. Given this, the appropriate model for capturing the immediate self-knowledge expressed by the content of 6 is some Fregean proposition consisting of the peculiar mode of presentation of Oedipus's expressed rule of employment of "I," roughly:

7. The individual employing a token of 6 (Oedipus) feels ashamed.

The meaningful employment of 7 relies on what Bermúdez has called the token-reflexive rule of the employment of the first-person pronoun:

8. If a person employs a token of "I," then he refers to himself by virtue of being the producer of that token. (Bermúdez 1998, p. 15)

Let us assume, just for the sake of argument, that Oedipus is not the Greek of the tragedy but some ordinary guy who calls his mom to excuse himself of his sins. Since his mother is not at home, the answering machine is automatically activated, and Oedipus utters sentence 6, recording it on the answering machine. Time goes by, and Oedipus forgets about it. He then returns home and checks the messages on the answering machine. However, when he listens to the messages from the answering machine, Oedipus does not recognize his own voice. What conclusion can we draw from this simple case?

First, Oedipus must assent to the content of sentence 7 (what he listens to when he hears the messages from the answering machine), provided only that he masters the rule of the employment of the first-person pronoun formulated in 8. Nonetheless, at the same time, he can deny the content of 6, modeled as a Russellian proposition, even though the contents of 5 and 6 are veritatively symmetrical. Even worse, as Oedipus does not recognize his own voice on the answering machine, even when he assents to the content of 7, he is not knowingly or reflexively self-referring. Reflexive self-reference requires the employer of the first personal pronoun to have knowledge of his own identity.

The problem is this: Since Oedipus was not born with the knowledge of the toke-reflexive rule that sentence 8 formulates, how could he have mastered it if he was not already self-conscious in the first place? Interestingly, Henrich responds to Tugendhat's criticism by claiming that "if we understand the word ("I") as an indexical word, the problem is eliminated" (1970, p. 49); that is, the problem is presupposed, not solved: we are back to Fichte's conundrum. As we will see in the next section, in order to master the rule of the first-person pronoun, Oedipus must be pre-linguistically aware of himself* as the employer of a token of "I." Tugendhat never took Henrich's answer seriously.

The following conclusions are imposing. Surely, Tugendhat is right when he claims that self-consciousness cannot be based on the traditional subject-object model (the notion of representation as "Vorstellung"). No one becomes self-conscious by occupying the subject and object poles of the representation ("Vorstellung") at the same time, i.e., by *identifying* oneself as the "represented" object of his intentional act as the "subject of representing," i.e., as the subject that performs the intentional act. An intellectual self-intuition in the sense of Fichte is out of the question here. But the invocation of token-reflexive rule 8 presupposes rather than solves the problem, for in sentence 8, self-identity is presupposed rather than explained. We are back to Fichte's puzzle: the use of tokens of "I" *presupposes* reflexive or knowing self-reference rather than explaining it.

Tugendhat's greatest mistake was to assume the original puzzle was linguistic rather than cognitive. Indeed, there is a revival of the theory of reflection in analytical philosophy under the label of higher-order theories of consciousness (see Rosenthal 1986, 1993, 2004, 2005, 2011, 2018; Carruthers 1989, 1996, 1999, 2000; Lycan 1987, 1996, 2001a, 2001b, 2004, and Gennaro 1996, 2004, 2005, 2006, 2012, 2015).

III. *Nonconceptual Self-consciousness*

The Heidelberg School and the phenomenological tradition both hold the view that through mastery of the token-reflective rule of the first-person pronoun, the individual is already "somehow familiar with himself" before ordinary reflexive or knowing self-reference takes place. The idea may sound plausible. Indeed, there does not seem to be any other possibility in sight. However, the idea of being "self-acquainted" is still nothing more than a metaphor! The question is: How can this metaphorical assertion be interpreted in concrete terms? In this section, I mention roads that I think we should avoid.

In his famous book, Bermúdez (1998) formulates another puzzle, namely the "paradox of self-consciousness:" in order to master the token-reflexive rule of the "I" according to which whosoever employs a token of the "I" knowingly refers to himself* presupposes that the individual employing a token of "I" is already conscious of himself* as the employer of the relevant token. Bermúdez claims that the only solution to his paradox is the postulation of "primitive nonconceptual forms of self-consciousness."

Bermúdez rightly rejects what he calls *The Conceptual Requirement Principle*, making room for the possibility of nonconceptual contents:

> ***The Conceptual Requirement Principle:*** *The range of contents that one may attribute to a creature is directly determined by the concepts that the creature possesses. (1998, p. 41)*

However, he is still attached to the linguistic dogma when he accepts Dummet's priority principle:

> ***The Priority Principle****: Conceptual abilities are constitutively linked with linguistic abilities in such a way that conceptual abilities cannot be possessed by nonlinguistic creatures. (1998, 42)*

We are back to the linguistic turn! I believe that a primitive, non-linguistic form of self-consciousness is required for the acquisition of the token-reflexive rule of the first-person pronoun "I." However, regardless of whether there are primitive nonconceptual forms of self-consciousness—a claim that Bermúdez himself recognizes as controversial—I do not see why the required non-linguistic self-consciousness must be nonconceptual.

On the phylogenetic scale, true perceptual systems appear in animal species long before the appearance of beliefs and propositional attitudes. Bees, frogs, pigeons, goldfish, and octopuses are, I suggest, good examples. Although they lack propositional attitudes, they have visual perceptual systems. The perceptual systems of some of these animals have been thoroughly studied. Scientific explanations of the discriminations, computations, and informational functions of the perceptual systems of lower animals individuate the representational content of visual states in part in terms of properties and relations in the animal's environment, properties and relations to which the animals maintain causal relations—both in sensory reception and in sensory activity. The best explanations for some of these low-level representational systems relate to the perception of physical objects in space and to the rudimentary spatial features of and between these objects. For example, the computations in the visual system of bees that relate to locating a hive operate on parameters that represent spatial positions and objects in those positions.

Nevertheless, there is overwhelming data to support the assumption that primates and other higher mammals have not only perceptions but also cognitions (propositional attitudes—beliefs, conceptualized desires, and intentions). The presence of beliefs presupposes the ability to reason, preserving propositional transitions between propositional attitudes, transitions that can be attributed as activities to the whole animal. Simple, logical, inductive, and meaningful reasoning is present in the mental activity of higher non-human animals. I also assume that primates and other higher mammals also have cognitions (propositional attitudes and even self-notions). A prey cannot think that a predator is coming towards it unless it has a notion of itself (a nonlinguistic form of self-consciousness).

There is a simpler cognitive solution to Bermúdez's paradox of self-consciousness within the mental file framework.[9] I assume that with the first brain maturation at the age of three to four months in the womb, the brain becomes "wired" and thus "acquainted with" its own body in a special, unique, and intimate way through the proprioceptive and kinesthetic channels so that there is a constant flow of information from whatever happens to the body to the brain. This is the non-metaphorical sense of Russell's "self-acquaintance." Now, with the completion of brain maturation at the age of three to four years, an "EGO-file" is subliminally created in the cortex with the precise function of storing proprioceptive and

kinaesthetic information about one's own body.[10] The crucial point is the following: Even if this "EGO-file" is pre-linguistic, it is not non-conceptual.[11] Therefore, Bermúdez's paradox can easily be solved without having to invoke the existence of nonconceptual forms of self-consciousness. With the help of this "EGO file," the creature first knowingly refers to itself as a prerequisite for mastering the token-reflexive rule of the "I."

The mental file framework also provides us with a much better explanation of what Tugendhat called "epistemic asymmetry" than Tugendhat's own account. According to Tugendhat's veridical symmetry, the propositions expressed in sentences 5 and 6 below are true under exactly the same conditions because they are both type-individuated as the same coarse-grained Russellian proposition: {Oedipus; feeling ashamed}.

5. Oedipus feels ashamed.
6. I think feels ashamed.

But the question is: why is the recognition of the same truth of the same coarse-grained proposition expressed in 5 and 6 "epistemically asymmetrical"? Mental files are "de re" modes of presentation of reference (see Recanati, 2012). Therefore, if we invoke mental files as neo-Fregean modes of presentation of the same individual, Oedipus, we split the content of 5 and 6 into two different fine-grained Fregean propositions ("Gedanken"): {Oedipus-file; the property of being ashamed}; {EGO-file; the property of being ashamed}. These fine-grained Fregean propositions are not the best expression of what 5 and 6 say. That is the Russellian proposition above. Be that as it may, they are the best expressions of Oedipus's state of mind in both contexts that account for the epistemic asymmetry.

Even if Bermúdez's postulation of primitive, nonconceptual forms of self-consciousness is not the best solution to his own "paradox of self-consciousness," the reader may wonder whether Bermúdez's postulation offers a solution to the old Fichtean conundrum. Unfortunately, this is not the case. Bermúdez's framework for his nonconceptual self-consciousness is the traditional subject-object model. The nonconceptual form of self-consciousness is a self-representation in which the same entity occupies both poles: that of "the representing subject" and that of "the represented object." Consequently, the Fichtean conundrum returns: in order to be able to refer to itself nonconceptually, the creature must already be self-conscious of its act of self-referring.

IV. *Pre-reflexive Self-consciousness*

As a way out of the dilemma, the phenomenologist postulates a pre-reflexive form of access to oneself. In such primary self-disclosure, one does not take oneself *as an object* either of one's inner perceptions or of one's thoughts. According to Sartre, for example, it is only the necessity of syntax that compels us to say that we are aware *of* our experiences or *of* ourselves. The basic claim is that one’s experiences and thoughts rely upon a peripheral awareness of oneself. When he focuses his attention on some cigarettes (Sartre's example), while he becomes transitively aware that they are twelve, he is also pre-reflexively aware that he is counting them. There is no infinite regress since, according to Sartre, "there is an immediate, noncognitive relation of the self to itself" (Sartre, 1956, p. 12).

The first thing that comes to mind is the question of why Sartre's pre-reflexive self-consciousness dispenses with self-identification altogether. Zahavi (2015) explains Sartre's metaphorical view based on Shoemaker's (1979) account of Wittgenstein's immunity to the error of misidentification through the use of the first-person pronoun. According to Shoemaker, there is no possibility of misidentification because there is no need for identification in the first place. The question is: Can this linguistic-analytic explanation of the special use of the first-person pronoun do justice to Sartre's phenomenological viewpoint? I do not think so. While Shoemaker speaks of a fundamental form of identification-free self-reference, Sartre seems to deny any self-reference when he talks about the putative pre-reflexive self-consciousness. Zahavi's proposal is a non-starter.

But let us suppose, just for the sake of argument, that Sartre's pre-reflexive self-consciousness is Shoemaker's identification-free self-reference. Is this a solution to the old Fichtean conundrum? The answer seems to be no! Sure, the phenomenological account may be able to stop the infinite regress by abandoning the old theory of reflection. But it cannot avoid Fichte's vicious circle. But why is this so? The phenomenological tradition remains stuck in the traditional representation framework (Tugendhat's subject-object model). This model comprises two poles: the subject pole—the representing subject—and the object pole—the represented object. Self-consciousness only occurs when the represented object is the same as the representing subject. But this is where Fichte's riddle comes into play again: the "representing subject" is always presupposed and never explained.

V. *Naturalizing Self-consciousness*

In this last section, I shall argue that we can only solve Fichte's puzzle by abandoning the traditional subject-object model completely. The proposal is not to replace it with a linguistic-analytical approach or with a phenomenological approach. The proposal is the naturalizing of the fundamental intentional relation of self-representation. Let me formulate my claim with a caveat: If there is a solution to the old Fichtean puzzle namely how to explain the emergence of self-representation without getting into an infinite regress or a vicious circle then naturalizing the concept of representation is the only way.

Let me begin with some well-known trivialities. According to Dretske (1988, 1995), two important factors determine the naturalization of the representational content of sensory experience. First, the "representational content" relies on a law-like connection between neural properties and distal properties of the environment: the instantiation of distal properties in the environment causes the instantiation of neurological properties in a law-like way. Conversely, due to this law-like causal connection, patterns of neural properties (as signs) convey information about what is happening in the external world (as a source), namely about the instantiation of distal properties of the environment.

For example, the size of a tree's annual rings correlates with factors such as the tree's age, soil richness, precipitation regime, and others. Similarly, the activation of specific nociceptive neurons in the parietal cortex in our species is strongly associated with various types of physical injury. Therefore, the activation of specific nociceptive neurons in the parietal cortex provides information about some physical injury.[12]

The second factor is what Dretske calls the "indicator function." Given the nomological correlations, some neurological

properties are then "recruited" (Dretske's metaphor of his 1988 book), either by natural selection or by learning, with the function of indicating (representing) the instantiation of the distal properties that are lawfully correlated with them.[13] When the representational system fulfills its function of indicating the instantiation of the distal property, the neurological representation is "accurate of" the particular instantiation of the property. However, if the system is not functioning properly or if it is not properly connected to the object, we have an "inaccurate representation."

For example, due to the nomological correlation between the activation of specific nociceptive neurons in the parietal cortex in our species and physical injury, the activation of these nociceptive neurons provides information about physical injury. The significance lies in the fact that such information is absolutely crucial for the survival of the species to which the creature belongs. Therefore, the naturalist assumes that the activation pattern of certain nociceptive neurons in the parietal cortex has been "recruited" by natural selection to indicate that the limbs or the body as a whole are injured.

Be that as it may, even his critics find Dretske's naturalization program audacious in many respects. However, one of the crucial achievements of the naturalization of representation is often overlooked in the literature: the departure from the traditional conception of representational content as "Vorstellung," in which the subject occupies both poles of representation. If Dretske's naturalization program is sound, it represents the first breakthrough in the old subject-object model. "Representation" no longer includes the two traditional poles of a "representation:" the "representing subject" and the "represented object." A complex biological system "represents" whatever physically lawfully correlates with it (a source) without a "representing subject" or an intentional act. Suppose a "neural pattern" is nomically correlated with a source and hence provides information about that source. Suppose we now assume that this flow of information is vital for the adaptation of the species to which the living being belongs. In that case, this neural pattern is "recruited" by selection with the biological function of indicating that source.

For example, when the activation pattern of nociceptive specific neurons in the parietal cortex has been "recruited" by natural selection to indicate that the limbs or the body as a whole are injured, there is no "representing subject" or "intentional act of representation." Similarly, when neurons in the cerebellum and brainstem are activated, neural patterns of two areas near the base of the brain are recruited to indicate posture, movement, and changes in balance, as well as knowledge of the position, weight, and resistance of objects in relation to the body (proprioception), without a subject representing anything. When the neurons in the posterior parietal cortex, as well as the primary motor cortex, are activated, neural patterns in both regions are also "recruited" to indicate the magnitude, direction, or weight of a movement (kinesthesia) without a "representing subject."

@

With the completion of brain maturation around the age of three to four years, all the self-relative information begins to be housed as a form of an autobiographic memory, which I call an "EGO file." For all we know, the posterior cingulate cortex, the anterior cingulate cortex, and the medial prefrontal cortex combine to provide humans with the ability to self-refer. However, it is the left dorsolateral prefrontal cortex and posterior cingulate cortex that are particularly involved in the memory of autobiographical information. The main assumption is that all the fleeting, dispersed manifold "self-relative"

information is "recruited" by natural selection as autobiographical memory, i.e., "memory for the events of one's life" (Conway & Rubin, 1993). In that moment, the subject starts to represent herself* as such without presupposing a subject. The fundamental form of self-reference is not consciously initiated by an "underlying subject." Instead, it emerges through the selfless "recruitment" of self-relative information housed as autobiographical memory.

The initial file functions as a temporary buffer, akin to a computer's buffer, lasting only a few hours or less. In this context, we can identify a nonconceptual form of self-awareness: The temporary nature of the buffer prevents its integration with representations of concepts within the cognitive system (as per Evans's generality constraint, Evans 1982). However, it is conceivable that the buffer, over time, evolves into a comprehensive EGO file hosting diverse self-information. At this juncture, well before language acquisition and the mastery of first-person token-reflexive pronouns, we can appropriately recognize the emergence of self-consciousness or knowing self-reference without a subject.

It is reasonable to speculate that the initial EGO file primarily contains information gleaned from proprioceptors, constituting what Tugendhat terms "epistemic immediate self-consciousness." As time progresses, the EGO file evolves into a comprehensive self-concept, capable of assimilating various types of self-information consistent with Evans's generality constraint (Evans, 1982). The EGO file eventually encompasses details acquired through the third-person pathway, such as the individual's age, height, and so forth, also falling under the category of "epistemic immediate self-consciousness," as posited by Tugendhat.

Returning to the puzzle of self-consciousness, the issue emerges from conceiving self-awareness within the subject-object framework, where the same individual occupies both subject and object poles of representation. The enigma arises because the "representing subject" must already be cognizant of herself as the agent conducting the representation. However, this puzzle is resolved as the fundamental form of self-reference emerges without a subject performing the act of representation. The EGO file is passively "recruited" by natural selection to indicate or represent the cognitive system itself.

VI. *Objections*

Let me draw a balance. Dretske's naturalistic account encounters several objections. Tyler Burge (2010), for example, claims that Dretske's (and others') program of naturalizing representation amounts to a trivialization of the rich notion of representation in cognitive science. If Dretske were right, thermometers would "represent" temperatures because their mercury column is nomologically related to temperatures, and we have given them the function of indicating temperature because of the lawlike connection between the mercury column and temperatures. Similarly, the magnetism of an anaerobic bacterium lawfully covaries with the deep, low-oxygen ionic regions of lakes, and it is reasonable to suppose that the magnetism was selected to indicate low-oxygen environments because of its benefit for the bacteria's survival. Yet, it is odd to claim that thermometers "represent" temperatures or that anaerobic bacteria "represent" low-oxygen environments.

Footnotes

[1] Following a convention that Castañeda created (Castañeda, 1966), I use the asterisk (*) to highlight the fact that the subject knows that he is referring to himself.

[2] Perry describes his situation as follows: "I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. But I seemed unable to catch up. Finally, it dawned on me. I was the shopper I was trying to catch." (Perry 1979, p. 33)

[3] Fichte had the great merit of being the first to recognize the problem. Since then, the continental tradition has wrestled with how to deal with the phenomenon of self-consciousness. To sidestep the conundrum, the phenomenological school has assumed a "pre-reflexive form of self-consciousness" as a necessary condition of reflexive self-consciousness (Sartre 1937, 1943, Introduction; Zahavi, 2005, 2007; Legrand, 2006; Kriegel, 2009). On the other hand, the so-called "Heidelberg School" has given up hope of solving this problem. Henrich calls the phenomenon a "riddle" (Henrich, 1966, p. 65) and portrays philosophical efforts to explain the phenomenon as utterly "hopeless." Cramer points out that self-consciousness confronts us with "an indisputable fact" whose explanation leads to difficulties that "seem almost insurmountable" (1974, p. 54). Similarly, Pothast (1971) describes the greatest challenge of cognitive self-reference as "insoluble."

[4] The entire passage is this:

> *"We become (...) conscious of our consciousness of our consciousness only by making the latter a second time into an object; thereby obtaining consciousness of our consciousness, and so on ad infinitum. In this way, however, our consciousness is not explained, or there is consequently no consciousness at all, if one assumes it to be a state of mind or an object and thus always presupposes a subject, but never finds it. The sophistry lies at the heart of all systems hitherto, including the Kantian." (Fichte, 1982, ii, p. 356).*

Henrich rephrases it as follows:

> *"It is not difficult to see that the reflection theory is* ***circular****: if we assume that reflection is an activity performed by a subject – and this assumption is hard to avoid – it is clear that reflections presuppose an "I" that is capable of initiating activity spontaneously,* ***for the "I" as a kind of quasi-act cannot become aware of its reflection only after the fact.*** *It must perform the reflection and be conscious of what it does at the same time as it does it." (1971, 11, emphasis added)*

However, Cramer certainly formulated the problem most clearly:

> *"But how can the subject know the 'she' in the reflection has herself as her own object? Apparently, only through the fact that the ego knows that she is identical with herself as her own object. Now, it is impossible to attribute this*

> *knowledge to reflection and to justify knowledge from it. Because for every act of reflection it is presupposed that I am already acquainted with myself, to know that the one with whom she is acquainted, when it takes herself as object, is identical to the one who is making the act of reflection turn back on itself. The theory, which wants to make the origin of self-consciousness understandable, therefore ends necessarily in a circle: that knowledge already must presuppose what it wants to explain it in the first place." (1974, p. 56.)*

[5] Rosenthal tries to avoid infinite regress by assuming that the higher-order thought that makes the lower-order thought conscious need not itself be conscious. He saw no problem if an unconscious subject of a higher-order thought could become self-conscious simply by identifying himself as the object of his unconscious higher-order thought. Let us, for the sake of argument, suppose that self-consciousness first when a first-order thought disposes the subject to identify himself as the object of his lower-order thought using his higher-order thought. The natural question is: How could the subject of a lower-order thought become self-conscious by means of higher-order thought if he was unconscious of himself as the subject performing the higher-order thought?

[6] In Fichte's words:

> *"The "I" posits itself absolutely, that is, without any mediation. It is at the same time subject and object. The "I" only comes into being through its self-positing – it is not a preexisting substance – rather, its essence in positing is to posit itself, it is one and the same thing; consequently, it is immediately conscious of itself." (1982, p. 357)*

[7] Tugendhat's entire complaint is as follows:

> *"The problem with the theory of reflection that Henrich identifies (…) rests on the assumption that we are analyzing something whose essence consists in the identity of knowing and what is known. For someone who does not acknowledge that the phenomenon of self-consciousness has or presupposes this structure, the difficulty does not exist. The difficulty, which is in fact insoluble, is only an outcome of the absurdity of the basic approach." (1979, p. 64)*

Tugendhat's solution involves a methodological re-orientation toward language:

> *"One asks oneself whether this problem disappears – or at least can be solved in any case – under the language-analytical view of epistemic self-consciousness, understood as that view that proceeds from the assumption that epistemic self-consciousness manifests itself in language, instead of relying on inner awareness." (1979, p. 54)*

[8] Regarding the use of the first-person pronoun, Tugendhat is quite clear:

> *"The word "I" refers to the speaker, that is, it stands for but does not identify this person. All other deictic*

> *expressions, as well as all other singular terms generally, refer by identifying, that is, they specify-in a direct or indirect way-how the intended individual object is to be located, and how it is to be distinguished from all others. By saying, for instance, "this beetle;' I identify it as this beetle here, to which I am pointing; by saying "tomorrow," I identify a certain day as that which comes after today, and the current day as that during which we are speaking, or I am speaking. But I myself am the one who speaks. I thereby identify myself, first, not for the sake of others, for when we do that, and we do sometimes, that has the same significance as when we extend our hand (he who answers "yes, it's me" on the phone identifies himself perhaps, though not through the word 'T' as much as through the sound of his voice). Second, I cannot identify myself with the word 'T' for my own sake, because I can only identify something by assigning it a position relative to myself-and so my body too has a place in the common general system of coordinates. That is why, in those "I" utterances in which I only set forth internal predicates of mine, I cannot be in error about the subject: in such utterances I cannot misidentify myself, not so much because, by saying "I;' I necessarily identify myself correctly, but rather because, by saying "I," I do not identify myself at all." (2016, pp. 8-9)*

[9] In connection with the referential use of definite descriptions (see Grice, 1969, pp. 140–144) or with identity statements (see Lockwood, 1971, pp. 208–211; Strawson, 1974, pp. 54–56), several philosophers in the late 1960s and early 1970s employed 'mental file' as a usual metaphor. This metaphor was subsequently adopted by various authors, including Evans (1973, p. 199 ff.; 1982, p. 276), Bach (1987, pp. 34–37), Devitt (1989, pp. 227–231), Forbes (1989, 1990, pp. 538–545), and Crimmins (1992, pp. 87–92). However, the most influential elaboration is due to John Perry (1980). Perry was certainly the first to transform the metaphor into an ingenious framework. With his 2012 book 'Mental Files,' Recanati systematized Perry's influential framework. Be that as it may, in this paper I still use the expression metaphorically.

[10] The distinction between channels is the reason for the difference between (1) and (2) in the Prolegomena.

The neuroanatomy involved in "self-representation" is quite complex. However, with the new neuroimaging technology, researchers have recently found several parts of the brain that are involved in "self-representation." These include the inferior parietal cortex, the anterior and posterior parts of the cingulate cortex, and the medial prefrontal cortex, especially its dorsal part that extends to the dorsolateral aspect (see Schmitz et al., 2004; Gusnard et al., 2001; Johnson et al., 2002; Keenan et al., 2000; Kelley et al., 2002).

[11] For one thing, the EGO-file meets Evan's generality constraint for concept attribution: EGO-file can be combined and recombined with quite different general properties and relations. For example, the same prey A that thinks that predator B is coming in its direction may also think, now as a predator, that its prey C is within A's reach. In other words, A can think that B is coming towards A but also think that A is coming towards C. If this is conceivable, A must possess a primitive pre-linguistic self-concept. The conclusion is the same as before: the 'linguistic turn' has got everything wrong again! See Evans 1982.

[12] Of course, there are several cases in which this is not the case: pain in phantom limbs; pains without any clear injury, etc.

[13] See Dretske 1988, p. 84, p. 99, p. 101, p. 102, p. 104, p. 113, p. 114, p. 124, p. 128, p. 147.

References

- Bermúdez, J.L. 1998. *The paradox of self-consciousness*. Cambridge: MIT Press.
- Carruthers, P. 1989. "Brute experience." *Journal of Philosophy*, 86: 258–269.
- ——, 1996. *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.
- ——, 1999. "Sympathy and subjectivity." *Australasian Journal of Philosophy*. 77: 465–482.
- ——, 2000. *Phenomenal Consciousness: a naturalistic theory.* Cambridge: Cambridge University Press.
- Cramer, K. 1974. Erlebnis, in: *H. Gadamer, Stuttgarter Hegel Tage*. Bonn.
- Fichte, J.G., 1794. *Grundlage der gesamten Wissenschaftslehre*. Jena und Leipzig.
- Gennaro, R. 1996. *Consciousness and Self-Consciousness.* Amsterdam: John Benjamins.
- ——, 2004. "Higher-order thoughts, animal consciousness, and misrepresentation." In R. Gennaro (ed.) 2004, pp. 45–66.
- ——, (ed.) 2004. *Higher-Order Theories of Consciousness*. Philadelphia: John Benjamins.
- ——, 2005. "The HOT theory of consciousness: between a rock and a hard place."*Journal of Consciousness Studies*, 12: 3–21.
- ——, 2006. "Between pure self-referentialism and the (extrinsic) HOT theory of consciousness." In Kriegel and Williford (ed.) 2006.
- ——, 2012. *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*. Cambridge, MA: MIT press.
- ——, (ed.) 2015. *Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness* Cambridge, MA: MIT press.
- Heidegger, M. 1989, *Die Grundprobleme der Phänomenologie*, Frankfurt a. Main.
- Henrich, D. 1966. *Fichtes ursprüngliches Einsicht*. Frankfurt a. M.
- — 1970: "Self-consciousness, a critical introduction to a theory."*Man and World* 4 (1): 3-28.
- — 2007: "Selbstsein und Bewusstsein." http://www.jp.philo.at/texte/HenrichD1.pdf.
- Husserl, E. 1973. *Zur Phänomenologie der Intersubjektivität III, Husserliana XV.* Den Haag Martinus Nijhoff.
- Lycan, W. 1987. *Consciousness.* Cambridge, MA: MIT Press.
- ——, 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.
- ——, 2001a. "Have we neglected phenomenal consciousness?"*Psyche*, 7. Available from the ASSC depository
- ——, 2001b. "A simple argument for a higher-order representation theory of consciousness."*Analysis*, 61: 3–4.
- ——, 2004. "The superiority of HOP to HOT." In R. Gennaro (ed.) 2004, pp. 93–114.
- Pothast, U. 1971. *Über einige Fragen der Selbstbeziehung*. Frankfurt am Main: Vittorio Klostermann.
- Rosenthal, D. 1986. "Two concepts of consciousness." *Philosophical Studies*, 49: 329–359.
- ——, 1993. "Thinking that one thinks." In Davies and Humphreys (eds) 1993.
- ——, 2004. "Varieties of higher-order theory." in R. Gennaro (ed.) 2004, pp. 17–44.

- ——, 2005. *Consciousness and Mind.* Oxford: Oxford University Press.
- ——, 2011. “Exaggerated reports: reply to Block.” *Analysis*, 71: 431–437.
- ——, 2018. “Misrepresentation and mental appearance.” *TransFormAcao*, 41: 49–74.
- Sartre, J.P., 1956. *The Transcendence of the Ego*. Trans. Forrest Williams and Robert Kirkpatrick. New York: Hill and Wang
- Shoemaker, S. 1996, the first-person perspective and other essays, New York, Cambridge University Press.
- Tugendhat, E. 1979. *Selbstbewusstsein und Selbstbestimmung. Sprachanalytische Interpretationen*. English translation: P. Stern. *Self-Consciousness and Self-Determination.* Cambridge, MA: MIT Press, 1986.
- Wittgenstein, 1958. *The Blue and Brown Books.* Oxford: Blackwell.