Forthcoming in *The Journal of Philosophy*

GUARANTEE AND REFLEXIVITY

Abstract. The rule account of self-conscious thought holds that a thought is self-conscious if and only if it contains a token of a concept-type that is governed by a reflexive rule. An account along these lines was discussed in the late 70s. Nevertheless, very few philosophers endorse it nowadays. I shall argue that this summary dismissal is partly unjustified. There is one version of the rule account that can explain a key epistemic property of self-conscious thoughts: GUARANTEE. Along the way, I will rebut a number of objections and introduce two constraints on how the reflexive rule is implemented.

Keywords. *I*-thoughts; self-consciousness; *I*-concept; guarantee.

According to a version of the tragedy, Oedipus once thought:

(1) *The solver of the Sphinx's riddle killed Laius.*

At the time of thinking, Oedipus ignored that he (himself) was the solver of the Sphinx's riddle. Sadly, a few years later, Tiresias was the bearer of bad news: "You're the solver of the Sphinx's riddle"—he said. So, Oedipus thought:

(2) *I killed Laius!*

(1) and (2) are about their thinker: Oedipus. Nevertheless, only (2) is a self-conscious thought. A theory of self-consciousness should identify the property that distinguishes (2) from (1). In addition, it should spell out the conditions under which thoughts like (2) instantiate that property.

According to the rule account, all it takes to have a self-conscious thought like (2) is to think a thought that contains a token of a concept-type that is governed by a reflexive rule (*RR*).[1] Unfortunately, the rule account has fallen into disgrace. Two reasons explain this situation. First, it has been argued that the rule account does not provide *sufficient conditions* for self-conscious thought.[2]

Second, several authors hold that the rule account does not even provide *necessary conditions* for self-conscious thought.[3]

Philosophers have been quick to devise alternative views. A conservative strategy is to supplement the reflexive rule with various forms of non-conceptual experience.[4] Another reaction is to posit a level of thought that is not governed by a reflexive rule.[5] A different approach grants that *RR* fixes the reference of the first person but insists that *RR* does not explain self-conscious thought.[6] A more radical response denies that the first person refers.[7] And other philosophers have tried to ground self-conscious thought in a primitive relation of self-acquaintance.[8]

It would go beyond the limits of this paper to examine each of these views. I will pursue a more modest goal. I will argue that this summary dismissal of the rule account is *partly* unjustified. There is a version of the rule account that can explain the epistemic property that distinguishes (2) from (1). This property is GUARANTEE. This version of the rule account can also rebut many influential objections. Thus, although I won't demonstrate that the rule account is true, my arguments should motivate the detractors of this view to take it more seriously.

Before we proceed, let me introduce some important assumptions. I shall assume that thoughts are episodic mental representations. In this framework, mental representations realize folk-psychological attitudes like beliefs or desires. I will remain neutral as to how the realization relation is to be understood. I will suppose that thoughts are complexes of concepts. On this view, concepts are constituents of mental representations that can be shared by different subjects and can participate in inference. One can construe thoughts either as representations in a Fodorian language of thought[9] or as internalized linguistic utterances.[10] What I shall say here is compatible with any of these views. Some authors characterize concepts as Fregean senses and thoughts as complexes of Fregean senses.[11] The current framework is certainly different from those approaches. Luckily, most of what I shall say here could be translated into a Fregean framework. I will use italics to mention concepts (*I*) and quotation marks to mention linguistic expressions ("I").

The article is structured as follows. I start with a characterization of GUARANTEE (Section I). Next, I introduce the rule account and dispel some common misunderstandings of it (Section II).

After that, I argue that the rule account provides necessary and sufficient conditions for GUARANTEE (Section III). Finally, I revisit Anscombe's circularity objection (Section IV) and the case of inserted thoughts (Section V).

I. Self-Consciousness and Guarantee

"Self-consciousness" and "self-conscious" are used as technical terms in philosophy. The former refers to a *capacity*, as when we say that some subjects have self-consciousness. The latter refers to a *property* of some mental states and events. Our focus is a specific property denoted by "self-conscious": GUARANTEE.[12]

Consider Oedipus' thought (2). This thought includes a token of the *I*-concept. It is widely held that all tokens of *I* are guaranteed to refer.[13] We can explicate this idea as follows:

DE JURE *REFLEXIVITY*. If a thinker, *S*, produces a token of *I*, the referent of that token is *S*.

DE JURE REFLEXIVITY distinguishes the way the *I*-concept refers from the way other concept-types refer. Suppose that Hume used the name *David* in soliloquy to think about himself. The fact that Hume used *David* to refer to himself is not part of how names refer. After all, names can refer to entities other than their users. The same holds for the concept-types underlying demonstratives and definite descriptions. Although they can be used to refer to their thinker, that is not part of how they refer. Let us follow Peacocke and say that the concepts underlying names, demonstratives, and definite descriptions can only be *de facto* reflexive; the *I*-concept is *de jure* reflexive.[14]

Self-conscious thoughts include a concept-type that is *de jure* reflexive. It is however unclear how the concept of *de jure* reflexivity relates to self-consciousness. An *epistemic* ingredient is missing. What is the missing ingredient?

Here is an influential view. If a thought is self-conscious, its thinker *must know* that *she herself* is the thinker of that thought.[15] Unfortunately, there are many controversies about the nature of

knowledge. We would ideally seek an account of self-conscious thoughts that has broad appeal, whatever one's conception of knowledge. Furthermore, the controversies about the nature of knowledge leave open an interesting possibility: there could be situations in which Oedipus' thought (2) realizes a merely true belief. However, (2) might still count as self-conscious in a philosophically interesting sense.[16]

These are not knockdown arguments against the knowledge approach. However, they give us some reason to explore an alternative. Here is a proposal. If a subject has a self-conscious thought, there are questions that she cannot raise concerning the referent of her own tokens of *I* while having that thought. Given that thinking a thought and asking a question take time, the phrase "while having that thought" should be understood as the lapse of time from the formation of the thought to the period of time during which the thought is held in the same stream of consciousness. So, if a subject has a self-conscious thought, she cannot ask *Does 'I' refer to me?* during that lapse of time. Let us use the phrase "questions of reference" to denote questions of this form. So, we have the erotetic criterion of GUARANTEE (from the Ancient Greek *erōtētikós*, which means "pertaining to questions"): If a thought has GUARANTEE, the thinker of that thought cannot raise questions of reference concerning her tokens of the *I*-concept while having that thought.[17]

The erotetic criterion hinges on the negation of a modal operator: "cannot." This modality is ambiguous. We will construe GUARANTEE as relying on two complementary interpretations of that modality: a normative reading and a constitutive reading.

According to a *normative reading*, if a subject's thought has GUARANTEE, that subject cannot *coherently* raise questions of reference while having that thought. Were the subject to raise those questions while having that thought, she would be irrational. Simultaneously thinking a thought that features a token of *I* and wondering whether *I* refers to me yields an incoherent combination of attitudes. In doing so, I am simultaneously treating the question of whether *I* is reflexive as "closed" (because I am employing a concept-type that leaves no room for non-reflexive uses) and as "open" (by asking whether *I* refers to me). And that seems irrational.[18]

According to a *constitutive reading*, if a subject's thought has GUARANTEE, a subject who raises questions of reference while having that thought does not count as having an *I*-thought. It is possible for some attitudes to display some form of irrationality without thereby being disqualified as instances of the relevant type. For example, it may be irrational to hold a belief on the basis of scant evidence. Yet, this does not disqualify the attitude as a token of the belief-type. Something different happens with Oedipus' thought (2). Were Oedipus to ask questions of reference while thinking (2), those questions would disqualify (2) as a token of an *I*-thought. Think about it. Suppose that Oedipus were to think (2) and asked: *Does 'I' refer to me?* In this scenario, most of us would conclude that, either *'I'* in (2) is not a token of the *I*-concept, or that Oedipus does not really master the *I*-concept.[19]

To be sure, a lot more could be said in relation to the erotetic criterion. However, the previous remarks suffice to provide a characterization of GUARANTEE:

GUARANTEE. A thought, *T*, has GUARANTEE if and only if:

(1) *T* has a token of a concept-type, *C*, that is *de jure* reflexive.

(2) A thinker of *T* cannot (coherently) raise questions of reference relative to *C* while having *T*.

GUARANTEE captures a key difference between (2) and (1). *The solver of the Sphinx's riddle* refers to Oedipus, the thinker of (1). However, *the solver of the Sphinx's riddle* is only *de facto* reflexive. Moreover, there are circumstances in which Oedipus thinks (1) and he can (coherently) raise questions of reference while thinking (1). He could (coherently) ask: *Does 'the solver of the Sphinx's riddle' refer to me?* This would happen if Oedipus ignores the identity: I = the solver of the Sphinx's riddle. By contrast, (2) features a token of a *de jure* reflexive concept. Moreover, when Oedipus thinks (2), he cannot (coherently) raise questions of reference while thinking (2). He cannot (coherently) ask: *Does 'I' refer to me?* Therefore, the erotetic criterion captures a key epistemic difference between (2) and (1).[20]

The rule account has often been dismissed. One reason for this summary dismissal is that "self-conscious" can be used to pick out different epistemic properties. In my view, the rule account cannot explain (without further conceptual tools) these other epistemic properties. I will argue, however, that the rule account can explain GUARANTEE, at least in the minimal sense of providing necessary and sufficient conditions for GUARANTEE. We can better appreciate the scope of the rule account by contrasting GUARANTEE with another epistemic property: IMMUNITY TO ERROR THROUGH MISIDENTIFICATION relative to the first person (IEM).

Some philosophers characterize self-conscious thoughts as thoughts that are IEM.[21] Given that "self-conscious" is a technical term, I have nothing against this approach. Nevertheless, IEM should be sharply distinguished from GUARANTEE.[22] We introduced GUARANTEE by comparing the epistemic profile of the *I*-concept with the epistemic profiles of the concepts underlying our uses of definite descriptions, proper names, and demonstratives. We noted that the *I*-concept is *de jure* reflexive. Moreover, all thoughts featuring the *I*-concept exclude some questions of reference. By contrast, IEM cannot be introduced by comparing different concept-types. We have to compare thoughts flanked by different predicative concepts based on different epistemic sources. For example, we can compare my thought *I have a headache* (based on a pain experience) with my thought *I am wearing a blue shirt* (based on a visual experience of a reflection in a mirror). The first thought excludes the following possibility: that the property λx(*x* has a headache) is instantiated and the bearer of that property is different from the referent of *I* (me). By contrast, the second thought leaves open the following possibility: that the property λx(*x* is wearing a blue shirt) is instantiated and the bearer of that property is different from the referent of *I* (me).[23]

It is easy to conflate GUARANTEE and IEM because both of them exclude identity mistakes. Nevertheless, GUARANTEE and IEM highlight two different ways in which identity mistakes can be excluded. GUARANTEE excludes the possibility that a thinker can (coherently) take the referent of her own token of *I* to be different from herself. IEM excludes a different possibility: that a thinker

can (coherently) take the bearer of a property that seems to be instantiated to be different from the referent of her own token of *I* (her).

An account of GUARANTEE may be part of an account of IEM. However, an account of GUARANTEE is not sufficient to explain IEM. An account of IEM should consider the way *I*-thoughts that are IEM are grounded in some epistemic sources. This goes beyond GUARANTEE, which also holds for *I*-thoughts that are not IEM. In our original example, Oedipus' thought (2) was based on Tiresias' testimony. Given that Tiresias might be talking of somebody else, the bearer of the property λx(*x* is the killer of Laius) could have been different from the referent of *I* (Oedipus). Yet, (2) still has an important epistemic property of self-conscious thoughts.

II. The Rule Account

Many writers have appealed to a reflexive rule to explain some epistemic property of self-conscious thoughts. In this section, I introduce (what I take to be) the strongest version of this family of views. This version is restricted to GUARANTEE and it relies on a specific formulation of the reflexive rule.

> *THE RULE ACCOUNT*. A thought, *T*, has GUARANTEE if and only if *T* contains a token of a concept-type that is governed by *RR*:

RR A token of *I* in a thinking stands for the subject of that thinking.

Let us work with a picture of subjects as possessors of mental states and events.[24] For our current purposes, we can remain neutral on the nature of subjects. The rule account has two components. The first component is *RR*. This rule states the relation that a subject must stand in to a token of *I* in order to be the referent of that token.[25] *RR* tells us that a token of *I* refers to the subject who stands in the thinking relation to that token. Second, the rule account relies on the concept of being

governed by a rule. To a first approximation, being governed by a rule contrasts with accidentally conforming to a rule. This contrast will allow us to block a pervasive objection (Sections III.2-III.5).

It has been argued that the rule account does not provide necessary and sufficient conditions for self-conscious thoughts (Introduction). Our formulation of the rule account challenges this view. A common tendency is to think that the rule account is meant to explain epistemic properties *other* than GUARANTEE. Our formulation preempts this interpretation.[26] In the remainder of this section, we will revisit other lines of criticism. More substantial objections will be discussed in subsequent sections.

OBJECTION 1. It might be objected that the rule account leaves out some important capacities that underlie self-conscious thoughts, including various ways of gaining self-concerning information (for example, proprioception, introspection), our ability to self-locate in space,[27] our awareness of other subjects,[28] and our consciousness of ourselves as agents.[29]

REPLY. It is true that the rule account does not mention any of these capacities. The rule account holds that those capacities are not relevant to explain GUARANTEE. However, it does not follow that those capacities are philosophically unimportant. They could be cited in other explanatory projects. For example, they could feature as *acquisition conditions*, that is, conditions that enable the acquisition of the capacity to think *I*-thoughts. They could be *background conditions*, conditions underlying the employment of first-person concepts. And they could be *sources of justification*, that is, epistemic sources for the self-ascription of properties.

OBJECTION 2. It is widely held that the rule account assumes that self-consciousness is language dependent. This overlooks the possibility that some forms of self-consciousness are non-linguistic.[30]

REPLY. It is true that the rule account has often been developed in the context of linguistic theories of the first-person pronoun.[31] Moreover, a commitment to the language dependence of self-conscious thought is implicit in some works that emphasize the centrality of a reflexive rule.[32] Notice, however, that our formulation of the rule account *does not entail* that all self-consciousness is language dependent. Our formulation is restricted to self-conscious *thoughts*. So, we can leave open

the question of whether *I*-thoughts are language dependent. What about non-conceptual forms of self-consciousness? Given that we have made use of *de jure* reflexive concepts in our characterization of GUARANTEE, our account certainly precludes the existence of non-conceptual representations endowed with GUARANTEE. However, "self-conscious" has been used in many other ways in the literature.[33] Crucially, our account does not preclude the existence of non-conceptual forms of self-consciousness *given other specifications of "self-conscious."*

OBJECTION 3. Some philosophers read the rule account as making the false prediction that all thoughts featuring a reflexive concept must be self-conscious.[34] However, this prediction is falsified by Anscombe's intriguing case of "A"-users.[35] In her thought experiment, speakers have an "A" stamped on the inside of their own wrists. Each speaker relies on the observation of her own "A" to report her own actions. Interestingly, "A" is reflexive: If a speaker, *S*, produces a token of "A," the referent of that token is *S*. However, a subject could mistake an "A" stamped on someone else's wrist for her own "A." Therefore, Anscombe concluded that utterances featuring tokens of "A" do not express self-conscious thoughts.

REPLY. Anscombe's thought experiment does not refute the rule account. As several authors have pointed out, the rule that governs "A" makes a non-eliminable reference to perception. A more accurate formulation of that rule would go as follows: *Any token of "A" refers to the subject on whose inside wrist she sees an "A" inscribed.* By contrast, *RR* makes no reference to perception.[36] The only way of refuting the rule account is to show that *RR* does not explain GUARANTEE. This is much harder, though.

We will consider more substantial objections as we proceed. I hope to have shown, however, that the rule account is more plausible than it is often taken to be. In the next section, I argue that the rule account *does* explain GUARANTEE.

III. Guarantee Explained

GUARANTEE has normative and constitutive dimensions (Section I). In this section, we will show how the rule account explains both of them.

III.1. Normativity. A thinker of (2) cannot *coherently* raise questions of reference while thinking (2). The rule account offers an elegant explanation of this fact.

When Oedipus thought (1), it was an accident that the referent of a token of *the solver of the Sphinx's riddle* was the *same as* the thinker of (1). Indeed, in many nearby situations, the referent of a token of *the solver of the Sphinx's riddle* is *not* the *same as* the thinker of (1). As an illustration, Jocasta could have easily thought (1). In that scenario, the referent of a token of *the solver of the Sphinx's riddle* would not be the *same as* the thinker of (1). By contrast, when Oedipus thought (2), it was not an accident that the referent of a token of *I* was the *same as* the thinker of (2). Indeed, there are no situations in which the referent of a token of *I* is *not* the *same as* the thinker of (2). Had Jocasta thought (2), the referent of that token of *I* would have been the *same as* the thinker of (2).

We have here an "asymmetry of accidentality." In (1), the thinker/referent identity is an accident. In (2), the thinker/referent identity is not an accident. If there are situations in which a token of a concept-type can refer to someone other than its thinker, then its thinker could have opportunities to *coherently* raise questions of reference in relation to tokens of that concept-type. For example, if Oedipus ignores that he (himself) is the solver of the Sphinx's riddle, it is not irrational for him to ask: *Does 'the solver of the Sphinx's riddle' refer to me?* By contrast, if there are no situations in which a token of a concept-type can refer to someone other than its thinker, it will be impossible for its thinker to *coherently* raise questions of reference in relation to tokens of that concept-type (while having the corresponding *I*-thought). So, the first step to offer an account of GUARANTEE is to identify a property of the concept-type *I* that explains why tokens of *I* must refer to their thinker.

We have seen that the *I*-concept requires that the referent of *all* its tokens be the same as its thinker. The *I*-concept is *de jure* reflexive (Section 1). By contrast, there is nothing in *the solver of the Sphinx's riddle* that requires that the referent of *all* its tokens be the same as its thinker. Thus, tokens

of *the solver of the Sphinx's riddle* can only be *de facto* reflexive. The rule account marks out the *de jure*/*de facto* difference by associating the *I*-concept with a rule that codifies the *de jure* reflexivity of the *I*-concept. This rule is *RR*: If a token of *I* in a thinking stands for the subject of that thinking, then Oedipus' act of thinking *I* guarantees that the referent of *I* is the same as Oedipus. Given *RR*, it is not an accident that the referent of *I* is the same as its thinker. Indeed, there are no situations in which a token of *I* refers to someone other than its thinker. Therefore, producing a token of a concept-type that is governed by *RR* will automatically deprive its thinker of any opportunity to *coherently* raise any question of reference concerning tokens of that concept-type (while having the corresponding *I*-thought). Were the subject to raise any question of reference while thinking a token of *I*, she would be irrational.

In sum, *RR* explains why raising questions of reference while thinking a token of *I* is irrational. Our next task is to explain why it is *constitutive* of *I*-thoughts that subjects do not raise questions of reference concerning their own tokens of the *I*-concept. We will introduce our explanation by reflecting on the relationship between *RR* and the cognitive role of the *I*-concept.

III.2. The Cognitive Role Problem. It has been argued that *RR* mischaracterizes the cognitive role of the *I*-concept. The argument goes as follows. If *RR* adequately characterizes the cognitive role of the *I*-concept, an *I*-thinker cannot raise questions of reference in relation to *RR*. But an *I*-thinker can easily raise questions of reference in relation to *RR*. So, *RR* does not adequately characterize the cognitive role of the *I*-concept.

John Campbell offers a telling version of this objection. On his view, *RR* is vulnerable to an "Open Question Argument":

> Moore famously said that any naturalistic reduction of goodness to a property F can be refuted by remarking that the question, "But is F good?" will always make sense.[37] Similarly, given any attempt to characterize the mode of presentation expressed by the first person, say as a mode of presentation X, the question, "But am I X?," will generally make sense.[38]

Suppose that, whenever a thinker produces a token of *I*, she explicitly represents *RR*: *A token of 'I' in a thinking stands for the subject of that thinking.* In this case, questions of reference will generally make sense: *Am I the subject of that thinking?* But those questions do not generally make sense.

Many philosophers have been moved by this type of argument. A natural way out is to deny that *RR* must figure as a description employed in thought. If *RR* does not figure as a description in thought, subjects won't have any opportunities to raise questions of reference about *RR*.[39]

A non-descriptivist formulation of the rule account should spell out the contribution of *RR* to the cognitive role of the *I*-concept. Otherwise, we would have failed to provide an explanation of GUARANTEE. To the best of my knowledge, we still lack a non-descriptivist formulation of the rule account. In the remainder of this section, I will fill this gap. First, I will argue that the rule account is *consistent* with the denial of the key assumption of Campbell's objection: subjects do not need to explicitly represent *RR*. Second, I will argue that the denial of the descriptivist assumption *follows* from the letter of the rule account plus some platitudes about being governed by a rule.

III.3. Producing Tokens of 'I' Our first task is to show that the rule account is *consistent* with the denial of the claim that *RR* figures as a description in thought. To this end, we will introduce two cognitive constraints on the implementation of *RR*: a "production constraint" (Section III.3) and a "consumption constraint" (Section III.4). After that, we will derive those constraints from the rule account (Section III.5).

Campbell does not explain how and when the Moorean question could arise. Therefore, we will need to make some assumptions. It may be helpful to situate *I*-thoughts in our mental life. Paradigmatic *I*-thoughts are responses to input conditions. Suppose that you have a headache. In this case, it is appropriate for you to think *I have a headache.* Imagine that you are looking at a city map of New York City with a red dot that says: "You are here." In this scenario, it may be appropriate for you to think *I'm close to Central Park*. One way of fleshing out Campbell's objection is to assume that the subject must explicitly represent *RR* whenever she moves from an input condition to a corresponding *I*-thought. Consider a "consultation model." On this model, self-referring is a complex process. Given an appropriate input condition, the subject entertains *RR*: *A*

token of 'I' in a thinking stands for the subject of that thinking. This representation leads our subject to identify a suitable subject that can stand in the thinking relation to a token of *I*. When the task is executed, the self-referring act is completed.

The consultation model faces obvious problems. To begin with, it could be argued that the consultation step cannot be executed, even in principle. It might be insisted that, in order to identify a subject that can stand in the thinking relation to a token of *I*, one must already think an *I*-thought.[40] But, if the consultation step is deemed necessary to follow *RR*, we are led to an infinite regress. Suppose now that the regress could be blocked, maybe by positing a mental demonstrative that refers to the thinker: *This is the subject of thinking*.[41] In this case, the consultation model would still yield the wrong prediction: *I*-thinkers should find it easy to raise questions of reference about their own tokens of *I*. A subject who is in the business of identifying the subject of a token of *I* could easily ask: *Does this token of 'I' refer to me? Am I the subject of thinking?* Nevertheless, *I*-thinkers do not find it easy to raise questions of reference. Therefore, the consultation model does not explain GUARANTEE.[42]

The good news is that the rule account neither mentions the explicit representation of *RR*, nor its consultation (Section II). Therefore, the rule account is *consistent* with a different picture. A natural alternative is to bypass the consultation step. The production constraint makes this idea explicit:

> *PRODUCTION CONSTRAINT*. When an input condition leads to an *I*-thought, the production of a token of *I* is underwritten by a basic, self-referring act.

Producing an *I*-thought involves two acts: a self-referring act and the act of self-ascribing a property. A satisfactory account of *I*-thoughts should say something about these two types of acts. However, only self-referring acts are relevant in this context. A self-referring act is an act of producing a token of the *I*-concept. Suppose now that self-referring is not a complex process that includes a "consultation step". Suppose further that there is no other activity mediating the production of the

self-referring act. Given these assumptions, the self-referring act is psychologically basic in a sense that is familiar from the philosophy of action.

Let us say that an act, *A*, is psychologically basic if and only if 1) *A* features in psychological explanations, but 2) *A* cannot be meaningfully factorized as the interplay of simpler, psychological acts. Baking a cake is not psychologically basic in this sense. Baking a cake is a complex act that can be meaningfully factorized as the interplay of simpler, psychological acts like breaking the eggs, mixing the ingredients, and so on. Baking a cake is therefore constituted by simpler, psychological acts. On pain of regress, baking a cake should bottom out in psychological acts that cannot be further factorized as the interplay of simpler, psychological acts. The suggestion is that, unlike baking a cake, self-referring acts cannot be meaningfully factorized as the interplay of simpler, psychological acts. So, following *RR* consists in performing a psychologically basic, self-referring act. This approach is *consistent* with the rule account. *RR* tells us that a token of *I* in a thinking stands for the subject of that thinking (Section II). Crucially, a basic, self-referring act *suffices* to put the subject in a thinking relation with a token of *I*.

We are now in a position to answer Campbell's objection. If an *I*-thinker can follow *RR* by performing a psychologically basic, self-referring act, she does not need to *find out* a suitable subject to stand in the thinking relation to a token of *I*. Therefore, there is no way in which our subject could possibly fail to find a referent or mistake the referent of *I* for someone else. In addition, the production of an *I*-thought won't generate any opportunities to raise questions of reference concerning tokens of *I*.[43]

III.4. Consuming Tokens of 'I.' One might think that the production constraint is sufficient for GUARANTEE. This view might gain some traction when we contrast the perspective of an *I*-thinker with the perspective of an external observer.[44] For an external observer, the *I*-thinker is always a perceptual object that might fail to be "there" or be mistaken for someone (or something) else. So, an external observer can have many opportunities to ask whether a token of *I* refers to *this* or *that* subject. The perspective of the *I*-thinker is different. From the perspective of the *I*-thinker, the

referent of *I* is always "there". There is no "distance" between the self-referring act and the referent. If I think *I*, the referent of *I* can only be *me*. Who else?

This reasoning is too hasty, though. To see why, let us consider an insightful remark from Anscombe:

> If you are a speaker who says "I," you do not find out what is saying "I." You do not for example look to see what apparatus the noise comes out of and assume that that is the sayer; or frame the hypothesis of something connected with it that is the sayer. If that were in question, you could doubt whether anything *was* saying "I."[45]

Although Anscombe's observation concerns the pronoun "I," it can be easily generalized to the *I*-concept. The consultation model construed self-referring as a complex process in which the prospective *I*-thinker had to find out the subject of her token of *I*. The production constraint removed the consultation step. Yet, nothing we have said so far excludes the possibility that the *I*-thinker performs a self-referring act but decides to interrupt the flow of thinking in order to find out what is thinking *I*. In this hypothetical scenario, it should be easy for the *I*-thinker to ask: *Does 'I' refer to me?* Were this question raised, her *I*-thought would lack GUARANTEE. Indeed, it would not plausibly count as an instance of an *I*-thought.[46]

What would it take for a subject not to raise questions of reference just after having produced an *I*-thought? Let us explore a minimal proposal: self-conscious *I*-thinkers must be wired in such a way that they do not raise questions of reference immediately after having produced their own *I*-thoughts. If we wanted to design a subject whose thoughts satisfy GUARANTEE, we should design her in such a way that she proceeds—in her subsequent doings—on the assumption that the referent of her own tokens of *I* is the same as the *I*-thinker. This yields a consumption constraint:

> *CONSUMPTION CONSTRAINT.* *I*-thinkers are disposed to immediately engage in activities that trade on the identity of the referent of their own tokens of *I* with the *I*-thinker.

The phrase "trade on identity" is best understood with an example. Suppose that you make the following inference:

Premise 1. Hesperus is F.

Premise 2. Hesperus is G.

Conclusion. So, Hesperus is both F and G.

When you performed the inference, you relied on the co-reference of the two tokens of *Hesperus*. In other words, you did not need an identity premise connecting the first token of *Hesperus* in premise 1 with the second token of *Hesperus* in premise 2:

Premise 3. Hesperus is Hesperus.

Now, suppose that the identity premise was needed. In this case, you would still need to rely on the co-reference of the two tokens of *Hesperus* in premise 3 with the tokens of *Hesperus* in premise 1 and premise 2 respectively. On pain of regress, there must be cases in which you simply rely on the co-reference of concept-tokens without representing an identity premise. When that happens, we can say that your transition "trades on the identity" of the two co-referential concept-tokens. In these cases, you do not ask yourself whether the two concept-tokens are co-referential. You simply take for granted that they are.[47]

The consumption constraint generalizes the concept of trading on identity in two ways. First, it applies it to the *I*-concept. Second, it extends it beyond the relation of co-reference. It does not only cover co-referential concept-tokens but also a token of the *I*-concept and subsequent activities performed by the referent of that token of the *I*-concept. (Please read "activities" in a sufficiently liberal way, including mental acts). Consider an example. Upon thinking *I have a headache,* an *I*-thinker can be led to massage her head or take a painkiller. In these cases, the *I*-thinker is proceeding on the assumption that the referent of her own token of *I* (in *I have a headache*) is the same as the *I*-thinker

who massages her head or takes a painkiller. By contrast, an Anscombean thinker who proceeds to find out what is thinking *I* is not disposed to immediately engage in activities that trade on the identity of the referent of her own token of *I* with the *I*-thinker. This unusual behavior will disqualify that thought as an *I*-thought.[48]

Figure 1 provides a schematic representation of the rule account:

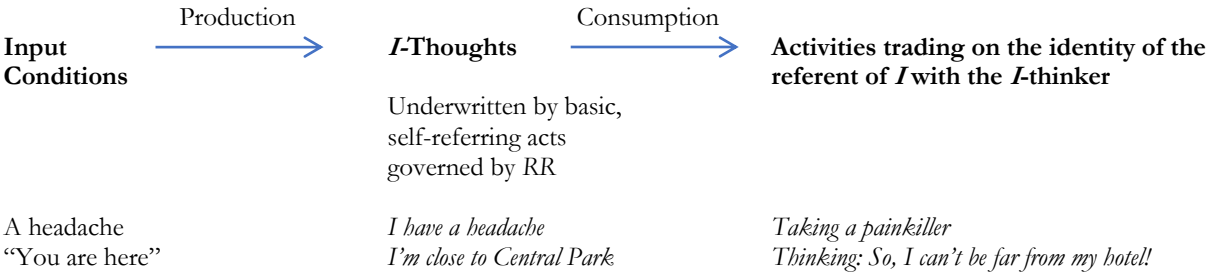

Figure 1. *The Rule Account*. The production of *I*-thoughts is partly underwritten by basic, self-referring acts governed by *RR*. *I*-thinkers do not need to find out the referent of *I* and they are disposed to immediately engage in activities that trade on the identity of the referent of their tokens of *I* with the *I*-thinker. This model offers subjects no opportunities to raise questions of reference while having *I*-thoughts. Therefore, it explains why *I*-thoughts satisfy GUARANTEE.

III.5. Being Governed by RR. It is natural to construe the production and the consumption constraints as conditions on the *implementation* of the *I*-concept. Some readers might wonder whether these implementation constraints are relevant to a *constitutive* account of self-conscious *I*-thoughts.

Notice that not all implementation conditions are irrelevant to address constitutive issues. The implementation conditions were introduced to provide an account of the cognitive role of the *I*-concept. Assuming that this cognitive role is constitutive of the *I*-concept, the two implementation conditions are part of a constitutive account of the *I*-concept. On this view, having a self-conscious *I*-thought partly consists in producing a basic self-referring act and having a disposition to immediately engage in activities that trade on the identity of the referent of *I* and the *I*-thinker. Chip away any of these constraints and the subject of the corresponding thought will not count as genuinely thinking an *I*-thought. In the remainder of this section, I will derive these two constraints

from the letter of the rule account plus some widespread assumptions about the concept of *being governed by a rule*. This should dispel the temptation to think of the previous constraints as *ad hoc*.

Recall the letter of the rule account: a thought, *T*, has GUARANTEE if and only if *T* contains a token of a concept-type that is governed by *RR*. The right-hand condition entails that no other rule (or combination of rules) is necessary for GUARANTEE. Otherwise, being governed by *RR* would not be sufficient for GUARANTEE. Models that posit the explicit representation and consultation of *RR* violate this requirement.

Suppose that *I*-thinkers were wired in such a way that they had to consult *RR* in the process of producing an *I*-thought. Our *I*-thinkers would have to find out the subject of the prospective token of *I* in a complex process of self-referring. In this scenario, *I*-thinkers would not simply perform a self-referring act; they would self-refer by doing other things: consulting *RR* and finding out a subject. Therefore, it would be a mistake to say that their use of the *I*-concept was governed *solely* by *RR*. Instead, their use of the *I*-concept would be governed by *RR* plus the sum of rules that govern the act of consulting *RR* and the process of finding out a subject. But this would go against the letter of the rule account. The rule account holds that producing a thought that contains a token of a concept-type that is governed by *RR* is both necessary and sufficient for that thought to have GUARANTEE (Section II).

When we take seriously the letter of the rule account, it seems to follow that nothing that falls short of a basic, self-referring act can implement *RR to the exclusion of implementing other rules*. *RR* tells us that *all* it takes for a subject to self-refer is to stand in a thinking relation to a token of *I*. Nothing else. The basic act of thinking *I* suffices to put the subject in a thinking relation to a token of *I*. If we held that self-referring is psychologically more complex, we would end up introducing further rules that govern the proper parts of the more complex psychological act.

What about the consumption constraint? It is widely accepted that *being governed by a rule* is different from *accidentally conforming to a rule*. Suppose that Oedipus happened to blink while he was thinking *I*. Therefore, Oedipus' token of *I* conformed to the blink-rule: *Blink while producing a token of I*. In this case, Oedipus' token of *I* would have *accidentally* conformed to the blink-rule. (According to

my sources, Oedipus did not have the slightest tendency to blink whenever he produced a token of *I*.)

The recent literature on rule following has sought to offer a principled distinction between accidentally conforming to a rule and being governed by a rule. This task is further complicated by the debate on meaning skepticism.[49] I won't engage in that debate. My aim is more modest. I will suggest that the consumption constraint is necessary to distinguish accidentally conforming to *RR* from being governed by *RR*. Even if we won't provide a full-fledged account of what it takes to be governed by *RR*, our argument will enable us to derive the consumption constraint from the letter of the rule account.

It is widely held that following a rule involves *at least* being disposed to proceed in certain ways.[50] It is not easy to state in the abstract, for all rules, what those ways are. However, it is easy to identify, for specific rules, some of those ways. Someone who adds 2+4=6 should be disposed to treat 2+5=7 as right. Were our subject not disposed to treat 2+5=7 as right, we would have a good reason to think that her first arithmetical operation is not governed by the successor function. So, having this disposition seems necessary to count her arithmetic activities as being governed by the rule of addition.

Something similar is true of *I*-thoughts. An Anscombean thinker who produces an *I*-thought but proceeds to find out what is thinking that thought is someone who is treating the referent of *I* as potentially different from the *I*-thinker. But *RR* precludes that the referent of *I* can be different from the *I*-thinker. Therefore, the Anscombean thinker is not proceeding in a way proper to someone whose tokens of *I* are governed by *RR*. This result strikes me as very intuitive. If we were to observe the behavior of the Anscombean thinker, we would conclude, either that she does not genuinely master *RR*, or that she is unwilling to commit herself to *RR*. So, in order to produce a token of *I* that is genuinely governed by *RR*, the *I*-thinker must have a disposition to immediately engage in activities that trade on the identity of the referent of her token of *I* with the *I*-thinker. *In this respect*, following *RR* is on a par with following other rules.[51]

III.6. Taking Stock. We have clarified the scope and theoretical commitments of the rule account. The rule account is not a theory of *anything* that philosophers have called "self-consciousness." It is an account of GUARANTEE. We have seen that the rule account provides necessary and sufficient conditions for GUARANTEE. Finally, the rule account does not say that *I*-thinkers must explicitly represent *RR*. Indeed, explicitly representing *RR* is inconsistent with the letter of *RR*. In the final sections, I revisit two influential objections to the rule account.

IV. The Circularity Objection

Anscombe famously argued that accounts of first-person reference by means of reflexive rules are circular. Although she was mostly concerned with the pronoun "I," her challenge can be generalized to the *I*-concept. Anscombe considered the following rule: "the word each one uses in speaking of himself."[52] The reflexive "himself" can be understood in two ways.[53] On one reading, it applies to cases of reflexive reference that are not self-conscious. Suppose that Oedipus utters:

(3) "The solver of the Sphinx's riddle killed Laius."

In this case, Oedipus produced a token of "the solver of the Sphinx's riddle" to speak of himself. Yet, he could (coherently) ask: *Does "the solver of the Sphinx's riddle" refer to me?* On the second reading, the reflexive exclusively applies to cases of reflexive reference that are self-conscious (what Anscombe calls the "peculiar indirect reflexive"). As an illustration, suppose that Oedipus utters:

(4) "I killed Laius."

In this case, Oedipus' token of "I" expresses self-consciousness. However, this second reading makes an ineliminable use of "I" in the elucidation of "himself." Therefore, the resulting account is

circular. Anscombe's diagnosis generalizes to *I*-thoughts. It suffices to replace (3) with (1) and (4) with (2).[54]

An initial response is that *RR* is different from Anscombe's reflexive rule. Given that *RR* does not include any occurrence of "him-" or "herself," it does not yield a circular explanation. Still, some readers might find our reply unsatisfactory. Indeed, some philosophers interpret Anscombe as introducing a more general, explanatory requirement: "The challenge is to provide an account of first-person reference that delivers genuine first-person reference without already assuming that very capacity in the explanation given."[55] The rule account posits basic, self-referring acts. Therefore, it does not satisfy Anscombe's explanatory requirement.

Let us use an analogy to evaluate the explanatory requirement. Suppose that we want to explain seeing. One option would be to posit a mechanism that produces a mental picture. If our hypothesis required an internal mechanism that "sees" the mental picture, the explanation would be circular. If seeing is not psychologically basic, this circularity would refute the proposed account of seeing. But notice the big "if." It would be implausible to hold that *any* psychologically interesting capacity must satisfy Anscombe's explanatory requirement. On pain of infinite regress, not all psychological capacities can be factorized as the interplay of more basic, psychological capacities.[56] In the absence of an argument to think the contrary, it is not unreasonable to hold that the capacity to self-refer is a basic, psychological capacity.

The explanatory requirement has motivated a research program seeking to investigate the phylogenetic and developmental antecedents of our capacity to think *I*-thoughts.[57] This research program is consistent with the claim that self-referring acts are psychologically basic. To illustrate, a pianist's complex capacity to play the piano may bottom out in basic capacities like the capacity to play individual notes. Yet, the development of the capacity to play the piano may still draw on other, background capacities.

V. Inserted Thoughts

Thought insertion is amongst the first rank symptoms of schizophrenia. In thought insertion, subjects have introspective access to some thoughts in their own stream of consciousness but insist that those thoughts are not their own. Some subjects also claim that those thoughts have been inserted into their minds by an external agent. One might wonder whether thought insertion challenges the rule account of GUARANTEE. I will concede that thought insertion is central to our understanding of various forms of non-conceptual self-consciousness. I will also grant that subjects who report inserted thoughts lack *fully* self-conscious *I*-thoughts. However, the phenomenon of thought insertion is consistent with the rule account of GUARANTEE. Before we reach these conclusions, we need to get a better understanding of this phenomenon.

V.1. Thought Insertion and Guarantee. As Christopher Frith points out, "[t]he lack of introspective data concerning thought insertion and related symptoms is surprising."[58] Therefore, most philosophical discussions of this phenomenon rely on a limited collection of verbal reports. Unfortunately, existing reports lack detail and background; it is often unclear whether they are literal transcriptions or examiners' notes from memory.[59] With these caveats, let us introduce a widely discussed example:

> Thoughts are put into my mind like 'Kill God.' It's just like my mind working, but it isn't. They come from this chap, Chris. They are his thoughts.[60]

Idealizing a bit, we can distinguish three types of thoughts in episodes of thought insertion:

> The disowned thought (*Kill God*).
> The disowning thought (*That is not my thought*).
> The explanatory thought (*That thought comes from this chap, Chris*).

The conjunction of these three thoughts is puzzling. However, many philosophers think that it is not incoherent. A dominant strategy has been to distinguish two strands in our concept of thought ownership. This distinction can be used to make sense of the combination of the disowned thought and the disowning thought.[61] Very roughly, a subject can be aware of having the thought *Kill God* in her own stream of consciousness without also being aware of being the author of that thought. More generally, the subject can have a sense of ownership for the disowned thought without also having a sense of authorship for that same thought. Therefore, in thought insertion, subjects deny authorship without also denying ownership.[62] There are many different ways of developing this strategy. However, these differences won't affect our discussion.[63]

In an engaging discussion, Christopher Peacocke draws some morals that are in tension with the rule account of GUARANTEE.[64] On his view, a schizophrenic subject could in principle experience a disowned thought of the form *I am F* and proceed to coherently ask: *Does 'I' refer to me? Or does it refer to an external agent?* Unfortunately, the rule account of GUARANTEE classifies this succession of thoughts as both incoherent and impossible.

I would like to resist the suggestion that the hypothetical schizophrenic subject has entertained a coherent and possible succession of thoughts. My starting intuition was that simultaneously thinking an *I*-thought and wondering whether *I* refers to me yields an incoherent combination of attitudes. In addition, I suggested that a subject who raises questions of reference while having an *I*-thought does not count as having an *I*-thought (Section I). Our hypothetical schizophrenic subject seems to lack a disposition to engage in those activities, for she treats the referent of *I* as (potentially) different from the *I*-thinker. Therefore, I would insist that she is either misusing the *I*-concept or is unwilling to commit herself to *RR* (Section III.5).

Thought insertion is consistent with the current analysis. I will grant that schizophrenic subjects have an impaired self-consciousness, *given other specifications of "self-consciousness."* Their condition may have sources in non-conceptual forms of experience that prevent those subjects from enjoying *fully* self-conscious *I*-thoughts. Nevertheless, the phenomenon of thought insertion does not challenge the rule account of GUARANTEE. To this end, I will first reject Peacocke's starting

assumption (Section V.2). After that, I will provide an interpretation of disowned *I*-thoughts that is consistent with those *I*-thoughts having GUARANTEE (Section V.3). At the end, I shall revisit Peacocke's view (Section V.4).

V.2. Peacocke's Starting Assumption. Peacocke assumes that *I*-thoughts can feature among the disowned thoughts reported by schizophrenic subjects.[65] Nevertheless, this assumption has been challenged.[66] In the original example, the disowned thought is in the imperative mood: *Kill God.* Imperative thoughts do not feature any token of the *I*-concept. Therefore, they are not in tension with the rule account of GUARANTEE. As indicated above, we lack a sufficient number of detailed reports of disowned thoughts. However, none of the reports I am familiar with justifies the assumption that *I*-thoughts can be disowned.

Some authors think of inserted thoughts as closely related to auditory verbal hallucinations in which subjects have experiences "of receiving a communication."[67] If we take the communication model seriously, it seems to exclude *I*-thoughts from the class of disowned thoughts. Imagine a subject who has the disowned thought: *I hate you*. Our subject could hardly understand the message if she were to take the token of *I* to refer to her herself. Instead, she should take the token of *I* to refer to someone else. More generally, messages that concern the receiver are not adequately "sent" via tokens of the first person, but rather via imperatives or representations featuring tokens of other pronouns, definite descriptions, and proper names.

V.3. Disowned I-Thoughts and Guarantee. Let us examine now the possibility that there are disowned *I-thoughts*. So, we could have the following combination of thoughts:

I am F (disowned thought).
But this is not my thought (disowning thought).
This thought comes from someone else (explanatory thought).

Let us concede, for the sake of the argument, that this combination of thoughts is not incoherent. This raises the question: Does this combination of thoughts constitute a counterexample to the rule

account of GUARANTEE? My answer is "no." Although these subjects have raised some questions, we lack reasons to believe that their questions concern the referent of their own token of *I* in the disowned thought.

To count as an Anscombean thinker, the subject should raise questions of reference concerning her own token of *I* in the disowned *I*-thought. In other words, our subject should treat the referent of her own token of *I* as (potentially) different from the *I*-thinker. However, it is not clear that the sort of interrogative behavior displayed by our putative schizophrenic subject concerns the referent of *I*. After all, a subject could raise all sorts of questions about a disowned *I*-thought without thereby wondering whether her token of *I* refers to her herself. Let us develop this point by considering Peacocke's treatment of inserted thoughts.

For Peacocke, the schizophrenic subject lacks an (apparent) action-awareness of thinking, that is, it does not seem to her that her thinking is something she is doing herself.[68] Whatever view of mental action one advocates, this view entails that authored thoughts differ from unbidden thoughts, that is, those thoughts that just occur to us, like tunes stuck in one's head or remembering an appointment just in time.[69] Thus, the schizophrenic subject could be understood as follows:

> *I am F* (because this token of *I* stands in the thinking relation to me). *However, this is not my thought* (because this thought was not an exercise of my own mental agency).

The action-awareness account is consistent with the rule account of GUARANTEE. To see why, it suffices to show that a subject who experiences an unbidden *I*-thought does not need to raise questions of reference concerning a token of *I* in her unbidden *I*-thought.

Consider a case adapted from Stephens and Graham.[70] Mary, a young mother concerned with her child's welfare and her maternal responsibilities, finds herself thinking *I am hurting my child.* Although her *I*-thought was unbidden, Mary does not need to raise the question: *Does 'I' refer to me?* Indeed, it would be hard to understand Mary's discomfort with her unbidden *I*-thought if she did not automatically take the token of *I* in her unbidden *I*-thought to refer to her herself. The unbidden

character of her *I*-thought does not prevent Mary from satisfying the consumption constraint. Upon thinking *I am hurting my child* Mary would probably hug her child in a protective manner or wonder: *Why am I thinking that?* Therefore, she would have engaged in some activities that trade on the identity of the referent of *I* and the *I*-thinker. Of course, the occurrence of the thought *I am hurting my child* will motivate Mary to raise many questions. Nevertheless, she does not need to raise questions of reference. Therefore, it is not unreasonable to assume that her token of *I* has GUARANTEE.

In sum, if the subject of an unbidden *I*-thought does not need to raise questions of reference concerning her own token of *I*, it is unclear why a schizophrenic subject who does not recognize her *I*-thought as an exercise of her own mental agency will raise questions of reference concerning a token of *I* in her disowned *I*-thought.

It might be objected that thought insertion and unbidden thoughts are different. Suppose that we define "thinking" as an activity that brings about thoughts.[71] Thus, a subject who denies authorship of her own thoughts is not just denying that her thought was not an exercise of her own mental agency. She is rather denying that she stands in the thinking relation to those thoughts. Notice, however, that this approach would presuppose the falsity of the action-awareness account. Moreover, it would turn the conjunction of the disowned thought and the disowning thought into an incoherent combination of thoughts. The subject would be simultaneously relying on the thinking relation between a token of *I* and herself in the disowned thought and denying that she stands in that thinking relation in the subsequent, disowning thought. However, we had assumed that the schizophrenic's combination of thoughts is coherent.

The same strategy can be applied to other interpretations of the experiential property that disowned thoughts lack. We can see whether lack of the relevant property is consistent with the subject's satisfaction of the consumption constraint. If, lacking that experiential property, the subject can still engage in activities that trade on the identity of the referent of *I* and the *I*-thinker, we have some reason to believe that those *I*-thoughts still have GUARANTEE.

V.4. Peacocke's Account Revisited. For Peacocke, the subject of a disowned thought "may know that occurrences of the first-person concept in thought express thoughts of a first-person type. Nevertheless, the first-person thoughts that occur to him during an experience of thought-insertion may not be ones of which he judges that the uses of the first person in them refer to him himself."[72] Our discussion suggests a different interpretation. It is unclear whether subjects can have disowned *I*-thoughts. If they can, the disowned *I*-thoughts could still enjoy an epistemic status different from knowledge that that thought is of the first-person type. I dubbed the property that characterizes those *I*-thoughts "GUARANTEE".

Our view contradicts the letter but not the spirit of Peacocke's view. Peacocke concedes that a subject can have knowledge of the referent of her own, unbidden *I*-thoughts.[73] So, he could grant that GUARANTEE is preserved in the absence of occurrent (apparent) action-awareness. However, he could still hold that some *I*-thoughts with GUARANTEE fail to exhibit the *full* self-consciousness proper to the *I*-thoughts that are accompanied by (apparent) action-awareness. The abnormal experience of thought insertion may be seen as motivating the subject to put forward extraordinary hypotheses about the "distal origins" of her own thoughts. In doing so, the subject will have engaged in a higher-order inquiry with the *I*-thought as its subject matter (footnote 19). However, those hypotheses need not contradict three important facts: 1) that she stands in a thinking relation to her own tokens of *I*, 2) that those tokens of *I* refer to her herself, and 3) that she is disposed to engage in activities that trade on the identity of the referent of *I* and the *I*-thinker (as the example of Mary illustrates). As the putative "recipient" of those *I*-thoughts, the schizophrenic subject can still count as having self-conscious *I*-thoughts in the minimal sense articulated by GUARANTEE.

VI. Conclusions

Very few philosophers endorse the rule account of self-conscious thought. I have argued that this attitude is partly unjustified. First, I have identified an epistemic property that the rule account can explain: GUARANTEE. Second, I have presented (what I take to be) the strongest version of the rule

account. Third, I have shown that this version of the rule account explains the normative and constitutive aspects of GUARANTEE. Fourth, I have argued that the rule account does not mischaracterize the cognitive role of the *I*-concept. Fifth, I have suggested that the rule account can respond to two influential objections.

SANTIAGO ECHEVERRI

Instituto de Investigaciones Filosóficas, UNAM

Acknowledgments. I am grateful to José Luis Bermúdez, Richard Dub, Miguel Ángel Sebastián, and an anonymous referee for their detailed comments on a previous draft of this paper. I am also grateful to Reinaldo Bernal, Romain Bourdoncle, David Chalmers, Jérôme Dokic, Béatrice Longuenesse, Ryan McElhaney, Ruth Millikan, David Papineau, Claudia Passos, François Recanati, and David Rosenthal for illuminating conversations on this topic. Work on this project was funded by a generous grant from the Swiss National Science Foundation (FNS P300P1_161061/1).

[1] See David Kaplan, "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford University Press, 1989), pp. 481–563, in particular pp. 533–4; John Perry, "Frege on Demonstratives," in his *The Problem of the Essential Indexical and Other Essays* (Stanford: CSLI, 2000), pp. 1-26; John Perry, "The Problem of the Essential Indexical," in *op. cit.*, pp. 27–44; Mark Sainsbury, "English Speakers Should Use 'I' to Refer to Themselves," in Anthony Hatzimoysis, ed., *Self-Knowledge* (New York: Oxford University Press, 2011), pp. 246–60.

[2] See Gertrude Elizabeth Margaret Anscombe, "The First Person," in Samuel Guttenplan, ed., *Mind and Language* (Oxford: Clarendon Press, 1975), pp. 45–65; José Luis Bermúdez, *The Paradox of Self-Consciousness* (Cambridge: MIT Press, 1998); Lucy O'Brien, *Self-Knowing Agents* (Oxford: Oxford University Press, 2007); Christopher Peacocke, *Truly Understood* (New York: Oxford University Press, 2008).

[3] Anscombe, *op. cit.*; Héctor-Neri Castañeda, "Reply to John Perry: Meaning, Belief, and Reference," in James E. Tomberlin, ed., *Agent, Language, and the Structure of the World* (Indianapolis: Hackett, 1983), pp. 313–28; Marie Guillot, "Thinking of Oneself as the Thinker: the Concept of Self and the Phenomenology of Intellection," *Philosophical Explorations*, XIX, 2 (August 2016): 138–60; François Recanati, *Direct Reference: From Language to Thought* (London: Blackwell, 1993).

[4] Bermúdez, *op. cit.*; O'Brien, *op. cit.*; Peacocke, *op. cit.*

[5] Perry, *Identity, Personal Identity, and the Self* (Indianapolis: Hackett, 2002); Recanati, *op. cit.*

[6] John Campbell, *Past, Space, and Self* (Cambridge: MIT Press, 1994); "Schizophrenia, the Space of Reasons and Thinking as a Motor Process," *The Monist*, LXXXII, 4 (October 1999): 609–25; "What is it to Know What 'I' Refers to?," *The Monist*, LXXXVII, 2 (April 2004): 206–18.

[7] Anscombe, *op. cit.*

[8] Annalisa Coliva, "The First Person: Error Through Misidentification, the Split Between Speaker's and Semantic Reference, and the Real Guarantee," this JOURNAL, C, 8 (August 2003): 416–31; "Stopping Points: 'I,' Immunity and the Real Guarantee," *Inquiry*, LX, 3 (Spring 2017): 233–52; Guillot, *op. cit.*; Saul A. Kripke, "The First Person," in his *Philosophical Troubles*, vol. 1 (New York: Oxford University Press, 2011), pp. 292–321; Sebastian Rödl, *Self-Consciousness* (Cambridge: Harvard University Press, 2007).

[9] Jerry A. Fodor, *The Language of Thought* (Cambridge: Harvard University Press, 1975).

[10] Daniel C. Dennett, *Consciousness Explained* (London: Penguin Books, 1991).

[11] Bermúdez, *op. cit.*; Gareth Evans, *The Varieties of Reference* (Oxford: Clarendon Press, 1982); Peacocke, *A Study of Concepts* (Cambridge: MIT Press, 1992).

[12] Our characterization of GUARANTEE builds on previous work by Coliva ("The First Person"; "Stopping Points); O'Brien (*op. cit.*); and Peacocke (*Truly Understood*; *The Mirror of the World: Subjects, Consciousness, and Self-Consciousness* (New York: Oxford University Press, 2014)). This characterization traces back to some remarks by Anscombe (*op. cit.*). Some differences will emerge as we proceed.

[13] But see Evans, *op. cit.*, p. 253.

[14] Peacocke, *The Mirror of the World*, p. 8.

[15] Anscombe, *op. cit.*, p. 47; Bermúdez, *op. cit.*, pp. 2-3; Campbell, *Past, Space, and Self*, p. 112ff.; Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard University Press, 1981), p. 79ff.; O'Brien, *op. cit.*, p. 56ff.; Perry, "Frege on Demonstratives," p. 15; Sainsbury, *op. cit.*, p. 246.

[16] Peacocke introduces the idea of a "fully self-conscious use of *I* in thought"—*Truly Understood*, *op. cit.*, p. 78. He defines it as a thought "in which the thinker knows that he is referring to himself, without drawing on any special information about the case." This leaves open the possibility that some uses of *I* are less than fully self-conscious. I will be concerned with one of these weaker forms of self-consciousness.

[17] For ease of exposition, I will often abbreviate "questions of reference concerning her tokens of the *I*-concept" to "questions of reference". What about subjects who lack the conceptual sophistication to ask questions of reference? Those subjects can still display patterns of *interrogative behavior* concerning their own tokens of the *I*-concept. I discuss one such pattern in Section III.4.

[18] Following Anscombe (*op. cit.*), O'Brien (*op. cit.*, p. 17) formulates a criterion of self-conscious thought in terms of doubt: "If 'I' refers, a comprehending user could not doubt that it refers and refers to the user herself." If doubt that *p* requires *dis*belief that *p* (or belief that *not-p*), the erotetic criterion is different. *Dis*believing *p* or believing *not-p* are not necessary to wonder whether *p*.

[19] Couldn't a philosophically oriented subject think an *I*-thought and ask a question of reference while having that thought? She could. However, this question would be part of a higher-order inquiry with the *I*-thought as its subject matter. See Campbell, *Past, Space, and Self*, p. 86.

[20] Coliva's concept of REAL GUARANTEE was a source of inspiration of GUARANTEE. There are two differences, though. First, Coliva appeals to the know-which construction: "the possession of the first-person concept guarantees that the subject knows which person that concept is a concept of"—"The First Person," p. 429. Given the context-sensitivity of know-which constructions, I prefer my formulation. Second, Coliva extends REAL GUARANTEE to demonstrative concepts—"Stopping Points." Given that demonstrative concepts are not *de jure* reflexive, GUARANTEE does not generalize to demonstrative concepts.

[21] Bermúdez, *op. cit.*, §1.2; Evans, *op. cit.*; Andy Hamilton, *The Self in Question: Memory, the Body and Self-Consciousness* (Basingstoke: Palgrave Macmillan, 2013); Perry, *Identity, Personal Identity, and the Self*; Simon Prosser and François Recanati, eds., *Immunity to Error through Misidentification* (Cambridge: Cambridge University Press, 2012); Sydney Shoemaker, "Self-Reference and Self-Awareness," this JOURNAL, LXV, 19 (October 1968): 555–67; Ludwig Wittgenstein, *The Blue and Brown Books* (Oxford: Blackwell, 1958).

[22] Several authors conflate GUARANTEE and IEM. See Anscombe, *op. cit.*, p. 51; Herman Cappelen and Josh Dever, *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person* (New York: Oxford University Press, 2013), pp. 133–5; Nozick, *op. cit.*, p. 90; Rödl, *op. cit.*, pp. 9–11, p. 124; Carol Rovane, "The Epistemology of First-Person Reference," this JOURNAL, LXXXIV, 3 (March 1987), pp. 147–67. Exceptions include Bermúdez, *op. cit.*, p. 9; Coliva, "The First Person"; "Stopping Points"; Hamilton, *op. cit.*, pp. 173–4; Peacocke, *Truly Understood*, p. 82; and Sainsbury, *op. cit.*, p. 259.

[23] This characterization may require some refinements. There is some disagreement about the nature of IEM, its scope, and its explanation. For example, we would need to distinguish *logical* and *de facto* IEM. See Prosser and Recanati, *op. cit.*; Max Seeger, "Immunity and Self-Awareness," *Philosophers' Imprint*, XV, 17 (July 2015): 1–19; Sydney Shoemaker, "Persons and Their Pasts," *American Philosophical Quarterly*, VII, 4 (October 1970): 269–85. In addition, we should generalize our characterization of IEM to some demonstrative thoughts. The next two paragraphs are restricted to IEM relative to the first person.

[24] Peacocke, *The Mirror of the World*, p. 38.

[25] Peacocke, *Truly Understood*, p. 57.

[26] Campbell argues that *RR* cannot explain self-consciousness—*Past, Space, and Self*, p. 112ff.; "Schizophrenia, the Space of Reasons and Thinking as a Motor Process," *op. cit.*, pp. 91–5. Most of his arguments purport to show that *RR* cannot explain IEM and various self-ascriptions. Similarly, Bermúdez construes the rule account as an attempt to explain IEM on the sole basis of *RR*—*op. cit.*, p. 10ff.

[27] Bermúdez, *op. cit.*; Evans, *op. cit.*; Peter F. Strawson, *The Bounds of Sense* (London: Methuen, 1966).

[28] George Herbert Mead, *Mind, Self, and Society* (Chicago: Chicago University Press, 2009).

[29] O'Brien, *op. cit.*; Peacocke, *Truly Understood.*

[30] This objection is implicit in Evans' and Recanati's critiques of Kaplan's and Perry's theories. See Gareth Evans, "Understanding Demonstratives," in his *Collected Papers* (Oxford: Clarendon Press, 1985), pp. 291–321, p. 321; Recanati, *op. cit.* pp. 67–8; Kaplan, *op. cit.*; Perry, "Frege on Demonstratives".

[31] Kaplan, *op. cit.*; Perry, "Frege on Demonstratives"; "The Problem of the Essential Indexical"; Sainsbury, *op. cit.*

[32] Campbell, *Past, Space, and Self*, p. 119ff.; Hamilton, *op. cit.*; Perry "The Problem of the Essential Indexical," p. 40.

[33] Bermúdez, *op. cit.*; Evans, *The Varieties of Reference*; Peacocke, *The Mirror of the World.*

[34] Anscombe, *op. cit.*; Nozick, *op. cit.*.

[35] Anscombe, *op. cit.*, p. 49.

[36] Campbell, *Past, Space, and Self*, p. 133; O'Brien, *op. cit.*, p. 57; Peacocke, *Truly Understood*, p. 84.

[37] George Edward Moore, *Principa Ethica* (Cambridge: Cambridge University Press, 1903).

[38] Campbell, "What is it to Know What 'I' Refers to?," p. 212. For similar objections, see Castañeda, *op. cit.*, p. 323; Guillot, *op. cit.*, pp. 141–2; Kripke, *op. cit.*, §1; Recanati, *op. cit.*, p. 71; and Rödl, *op. cit.*, pp. 1–2, p. 180.

[39] Peacocke, *The Mirror of the World*, p. 84; David Rosenthal, "Awareness and Identification of Self," in JeeLoo Liu and John Perry, eds., *Consciousness and the Self* (Cambridge: Cambridge University Press, 2012), pp. 22–50. The assumption that *RR* is a description employed in thought is explicit in Campbell's treatment. See also Guillot, *op. cit.*, p. 140; Kripke, *op. cit.*, pp. 300–1, p. 304; Recanati, *op. cit.*, p. 72; and Rovane, *op. cit.*, pp. 86–8, p. 91.

[40] Bermúdez, *op. cit.*, p. 17; Coliva, "The First Person," p. 430; Kripke, *op. cit.*, §1; Recanati, *op. cit.*, pp. 71–2.

[41] Kaplan, *op. cit.*, pp. 534-5; Rovane, *op. cit.*, pp. 87-8.

[42] Rovane (*op. cit.*) posits an infallible self-identification ability. I find this view hard to swallow. Moreover, the possession of an infallible self-identification ability is insufficient to prevent a subject from raising questions of reference. After all, one can have an infallible ability without being aware that one has it. What is needed is an account in which subjects are aware of their possession of an infallible self-identification ability. Alas, I am not aware of having it.

[43] Some philosophers have introduced self-acquaintance as an alternative to the rule account. Coliva, "The First Person," p. 430; Guillot, *op. cit.*; Kripke, *op. cit.*, p. 301; Recanati, *op. cit.*, p. 72ff.; Rödl, *op. cit.* Those philosophers might protest that our production constraint comes very close to an acquaintance view.

It is worth noting that "acquaintance" can be used in different ways. If it denotes *any* non-descriptivist reference fixing mechanism, then our production constraint introduces a form of acquaintance. Notice, however, that the production constraint leaves out many of the properties traditionally associated with acquaintance: 1) it is compatible with Hume's elusiveness thesis: even if the self is not manifest in introspection, a subject can stand in a thinking relation to a token of *I*; it suffices that she thinks *I*; 2) acquaintance is often construed as a form of experience that reveals the essence of things; the production constraint does not have this strong implication; 3) acquaintance is often construed as a standing relation that we all bear to ourselves and/or our experiences; the thinking relation is rather an episodic relation that we bear to our own tokens of *I* when we are thinking them.

[44] Tomis Kapitan, "Indexical Identification: A Perspectival Account," *Philosophical Psychology*, XIV, 3 (Summer 2001), pp. 293–312.

[45] Anscombe, *op. cit.*, p. 56.

[46] The attempt to find out what is thinking *I* is an example of interrogative behavior applied to tokens of the *I*-concept. See footnote 17.

[47] Campbell, *Past, Space, and Self*, p. 74ff.

[48] Several philosophers have argued that indexicals are immediately relevant for action guidance. See Héctor-Neri Castañeda, "'He*': A Study in the Logic of Self-Consciousness," *Ratio*, VIII (1966): 130–57; Perry, "Frege on Demonstratives"; "The Problem of the Essential Indexical."

A subject who is disposed to run away upon thinking *That bear is about to attack me* satisfies the consumption constraint. Still, the consumption constraint is compatible with views according to which the immediate precursors of action are not indexical thoughts but rather non-conceptual *de se* representations (Bermúdez, *op. cit.*; Peacocke, *The Mirror of the World*) or representations with unarticulated *de se* constituents (Campbell, "What is it to Know What 'I' Refers to?"; Perry, "Thought without Representation," *Proceedings of the Aristotelian Society, Supplementary Volumes*, LX, 1 (July 1986): 137–52).

[49] Saul A. Kripke, *Wittgenstein on Rules and Private Language* (Oxford: Blackwell, 1982).

[50] Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Blackwell, 1953), §185.

[51] There is a key difference, though. I can make a mistake while following the rule of addition. But it does not seem possible to make a mistake while following *RR*. A Wittgensteinian objection says that, in the absence of the possibility of being wrong, one cannot be right either. This objection is inconclusive. We could see *RR* as a limiting case of following a rule. Moreover, we could insist that tokens of *I* can be embedded in complex thoughts that can display various types of error.

[52] Anscombe, *op. cit.*, p. 47.

[53] Castañeda, "'He*'".

[54] Bermúdez, *op. cit.*, §1.4; Hamilton, *op. cit.*, §1.3; Nozick, *op. cit.*, pp. 80–1; O'Brien, *op. cit.*, p. 56.

[55] O'Brien, *op. cit.*, p. 6; see also Bermúdez, *op. cit.*, §1.4.

[56] Jennifer Hornsby, *Actions* (London: Routledge & Kegan Paul, 1980).

[57] Bermúdez, *op. cit.*

[58] Christopher D. Frith, "Comments on Shaun Gallagher," *Psychopathology*, XXXVII, 1 (January 2004): 20-2, p. 22.

[59] Gottfried Vosgerau and Martin Voss, "Authorship and Control Over Thoughts," *Mind and Language*, XXIX, 5 (November 2014): 534–65.

[60] Frith, *The Cognitive Neuropsychology of Schizophrenia* (Hove: Lawrence Erlbaum, 1992), p. 66. For other examples, see Christoph Hoerl, "On Thought Insertion," *Philosophy, Psychiatry & Psychology*, VIII, 2/3 (June-September 2001): 189–200; Karl Jaspers, *General Psychopathology* (Manchester: Manchester University Press, 1963); C. S. Mellor, "First Rank Symptoms of Schizophrenia," *British Journal of Psychiatry*, CXVII, 536 (July 1970): 15–23.

[61] An elucidation of explanatory thoughts requires additional theoretical resources that go beyond the two strands in the concept of thought ownership. I will put explanatory thoughts to one side.

[62] Campbell, "Schizophrenia, the Space of Reasons and Thinking as a Motor Process"; Shaun Gallagher, "Self-Reference and Schizophrenia: A Cognitive Model of Immunity to Error Through Misidentification," in Dan Zahavi, ed., *Exploring the Self: Philosophical and Psychopathological Perspectives on Self-Experience* (Amsterdam: John Benjamins, 2000), pp. 203-39; G. Lynn Stephens and George Graham, *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts* (Cambridge: MIT Press, 2000); Hoerl, *op. cit.*; Peacocke, *Truly Understood*, §7.8.

[63] Some philosophers hold that subjects do not even retain the sense of ownership. See Alexandre Billon, "Does Consciousness Entail Subjectivity? The Puzzle of Thought Insertion," *Philosophical Psychology*, XXVI, 2 (Spring 2013): 291–314; Coliva, "Thought Insertion and Immunity to Error through Misidentification," *Philosophy, Psychiatry & Psychology*, IX, 1 (March 2002): 27–34; Lisa Bortolotti and Matthew Broome, "A Role for Ownership and Authorship in the Analysis of Thought Insertion," *Phenomenology and the Cognitive Sciences*, VIII, 2 (June 2009): 205–24. For reasons of space, I cannot discuss these views here.

[64] Peacocke, *Truly Understood*, pp. 89-90.

[65] See also Campbell, "Schizophrenia, the Space of Reasons and Thinking as a Motor Process," p. 621.

[66] Coliva, "Thought Insertion and Immunity to Error through Misidentification," p. 30; Seeger, *op. cit.*, p. 4. Relatedly, O'Brien (*op. cit.*, p. 70) speculates that disowned thoughts are demonstrative thoughts.

[67] Frith, *The Cognitive Neuropsychology of Schizophrenia*, p. 73; Peter Langland-Hassam, "Fractured Phenomenologies: Thought Insertion, Inner Speech, and the Puzzle of Extraneity," *Mind and Language*, XXIII, 4 (September 2008): 369–401, p. 373.

[68] Peacocke, *Truly Understood*, pp. 276–7.

[69] Peacocke, *Truly Understood*, p. 249.

[70] Stephens and Graham, *op. cit.*, p. 173.

[71] Vosgerau and Voss, p. 542.

[72] Peacocke, *Truly Understood*, *op. cit.*, p. 90.

[73] Peacocke, *Truly Understood*, *op. cit.*, pp. 90–1.