The Philosophical Quarterly Vol. 0, No. 0 *2024*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqae133

The *Real Guarantee* in *De Se* thought: How to characterize it?

Manuel García-Carpintero

LOGOS/BIAP-Departament de Filosofia, Universitat de Barcelona, Spain

Castañeda, Perry and Lewis argued that, among singular thoughts in general, thoughts about oneself 'as oneself'—first-personal thoughts, which Lewis aptly called de se—have a distinctive character that traditional views of contents cannot characterize. Drawing on Anscombe, Annalisa Coliva has argued that a feature she calls Real Guarantee marks apart de se thoughts—as opposed to others including Immunity to Error through Misidentification that have been proposed for that role. I'll argue that, while her work points to a truly distinguishing feature of the de se, we need an account of the notion other than hers and Echeverri's recent development of it. I'll offer an alternative, drawing on Léa Salje's work. Finally, I'll briefly outline how, thus understood, the Real Guarantee feature could be adequately explained by theories of de se thoughts like the one I favour, even if it is an open option to just explain it away.

Keywords: first-personal reference; *de se* attitudes; subjectivity; self-knowledge; immunity to error through misidentification.

I. Preamble: the *Real Guarantee* and the *De Se*

Castañeda (1966, 1968), Perry (1979) and Lewis (1979) argued that, among singular thoughts in general, thoughts about oneself *as oneself*—first-personal thoughts, which Lewis aptly called *de se*—have a distinctive indexical character that traditional views of contents overlook. Drawing on some remarks by Anscombe, in a series of papers Annalisa Coliva (2003, 2006, 2012, 2017) argues that a feature she calls *Real Guarantee* ('RG' henceforth) is a better candidate to distinguish *de se* thoughts from others than *Immunity to Error through*

Correspondence to: Manuel García-Carpintero, m.garciacarpintero@ub.edu

Misidentification ('IEM'), which Anscombe (1975, 57) also discusses, and Evans (1982) and others mention in that regard. This is what Anscombe says in the most significant passage:

> 'I' – if it makes a reference, if, that is, its mode of meaning is that it is supposed to make a reference – is secure against reference-failure. Just thinking 'I …' guarantees not only the existence but the presence of its referent. It guarantees the existence because it guarantees the presence, which is presence to consciousness. […] here 'presence to consciousness' means physical or real presence, not just that one is thinking of the thing. For if the thinking did not guarantee the presence, the existence of the referent could be doubted. For the same reason, if 'I' is a name it cannot be an empty name. I's existence is existence in the thinking of the thought expressed by 'I …' (Anscombe 1975: 55)

Anscombe's point is that self-reference, if real, would be guaranteed, because one is 'really present' in the act. Anscombe argues that a Cartesian ontology of the self is thereby vindicated *if* the condition is granted—i.e., that 'I''s 'mode of meaning' is that 'it is supposed to have a referent'.[1] She famously rejects that it is. Sainsbury (2011: 260) and Coliva (2012: 32–40) show that her argument for this is unconvincing.[2] At the very least, the 'no-reference' view that Anscombe appears to defend—on which 'I' is taken to work like expletive 'it' in 'it is raining'—is costly. Doyle (2016), Wiseman (2017), Haddock (2019) and Stainton (2019) offer interpretations of Anscombe 'no-reference' view which don't assimilate 'I' to expletive 'it'; but I'll argue below that they fail to identify an intermediate position (fn. 29).

I believe that Coliva's discussions reveal a truly distinguishing feature of the *de se*, which doesn't necessarily compel a Cartesian analysis. She (cf., e.g., Coliva 2017: 239) shows that some *de se* thoughts with RG nonetheless lack IEM—the feature that Evans (1982) and others consider discriminating. IEM is the impossibility for a judgement of the form 'a is F' of an 'error through misidentification', which occurs just when the source object (i.e. the object from which the predication information derives, if any) is different from the target object (i.e. the object to which the predicate is applied, if any).[3] Moreover, not only *de se* thoughts exhibit IEM. Thus—to confront sceptics like Millikan (1990), Cappelen and Dever (2013) or Magidor (2015)—RG intuitively has the potential to illuminatingly distinguish thoughts about oneself *as oneself* (those that

[1]Anscombe's view has strong Kantian overtones that Salje (2020: 744–5) aptly glosses.

[2]Salje (2020) makes a powerful case that Anscombe's Lichtenbergian train of thought can be characterized as a *cognitive illusion*—a Wittgensteinian philosophical blunder, we may say. But see Wiseman (2019) for compelling considerations to be careful in interpreting Wittgenstein and, by extension, Anscombe on these issues.

[3]This slightly modifies Seeger's (2015a: 2) account, based on Prosser (2012: 161–2); see also Wright (1998: 19) and Campbell (1999: 89) for similar notions. McGlynn (2021: 2305) has a more precise definition; see also Echeverri (2020: 480), Morgan (2019: 446–7) and McGlynn (2021: 2309–10) for discussion.

deploy the indexical self-concept SELF$_I$) from those that one has about oneself by deploying instead a proper-name-like concept NN (Coliva 2017: 237).[4]

Anscombe describes RG in essentially pre-theoretical terms. It ultimately just is, I think, the feature that makes Descartes' *cogito* intuitively compelling: that properly first-personal means of reference (if indeed they are such means) have their reference guaranteed because the referent is 'present' in the very act of reference. It would be very good to have a clear-cut characterization of this feature compatible with as many philosophical views as possible and as close as possible to folk pre-theoretical notions, because then it could be plausibly argued that it is a crucial datum for such philosophical undertakings to account for, or, alternatively, to explain it away. Coliva does offer one; but I don't think it characterizes RG well enough to achieve these goals. Coliva appeals to a knowing-wh criterion which, as shown by Boër & Lycan (1986), is context-dependent in ways that, as I'll show, makes it problematic to use it for these purposes. I'll articulate a dilemma (Section II): given Coliva's characterization, either clear cases of *de se* thoughts fail to have RG—if Coliva's knowing-wh criterion is allowed to be contextually interpreted without any restriction (first horn); or (second horn), if—as I think she intends—it is restricted,[5] it fails to distinguish first-personal thoughts deploying SELF$_I$ from those deploying NN. In recent work, Echeverri (2020: 478 fn.) alludes to the problem I'll raise,[6] and he provides an alternative *erotetic* account of RG (*ibid.*, 477) that echoes other suggestions by Coliva. Alas, it fares no better than Coliva's, as I'll show (Section III) by plotting a related dilemma.

I should make it clear that my discussion is meant to offer friendly amendments to Coliva's and Echeverri's accounts; my views on these issues are quite close to theirs, I think. The alternative proposal I'll make in Section IV to characterize RG is intended to elaborate and clarify theirs, so that we can rely on RG for the shared goal of critically examining philosophical accounts of self-knowledge and self-awareness. I'll argue there that, despite her own qualms, Salje (2020) provides a better elaboration of the suggestions in Anscombe's quote.

I'll close (Sections V and VI) by outlining how a non-Cartesian view on first-personal thought can explain RG.[7] These sections are not meant as a full-fledged defence of my account, which cannot be done in a piece like this to any satisfactory measure. They are just meant to support my main

[4] Cf. Salje (2019) and Echeverri (2020, 2021) on the need to posit, and distinguish, such self-concepts. In Recanati's (2016) 'mental file' framework, NN would be an 'encyclopedia entry'.

[5] Coliva doesn't elaborate on what the restriction should be; I'll offer one consistent with her remarks that I think apt: a restriction to semantic, reference-fixing identifying information (Section II).

[6] Echeverri doesn't explain why it is problematic, as I'll do; see below, Sections II and III.

[7] Coliva (2012: 37–9; 2017: 241) countenances a view close to the one I favour. She rejects it, but its central features are, I take it, very close to her own stance on these matters. It also shares important features with Sainsbury's (2011) and Echeverri's (2020, 2021) views.

claim, to wit: that RG is a pre-theoretical notion that good accounts of the *de se* should either honour or dismiss with good reasons. For this it should, and can, be characterized independently of such theoretical views. The point that some accounts in the literature accomplish this is meant to support the characterization that I'll pluck from Salje, by showing that, thus understood, RG specifies a mark of first-personal reference that apt philosophical accounts can explain. Sceptics however may explain it away, and there may well be alternative accounts equally capable of explaining it. The one outlined here would thus need more elaboration, at the very least.

II. Coliva's *Knowing-wh* characterization of the *Real Guarantee*

This is the way Coliva (2017: 235) sums up the contrast that RG sets up between her thinking thoughts about herself using a proper-name-like concept *AC*, and using instead the first-person concept: 'while when I entertain the proposition "AC is F", I may devise scenarios in which I go on sensibly to ask "Ok, AC is F, but which person is AC?", once I home in an I-content, then I can no longer sensibly ask "OK, I am F, but which person is I?"'.[8] But in fact I can sensibly ask questions with this very form: 'I am in this picture, but which person is I?' Or, for one more case: I am about to marry Alex. I find out that there is an old prophecy, that Alex will marry either The Prince, and they would have a happy life, or The Beggar, and they would have a life of misery. I say: 'I'll marry Alex, but which person is I?' Of course, when asking such questions, I still know very well (by the standard in Anscombe's quote at the outset) that it is me whom I am referring to; I know the person I lack the resources to identify among those in the picture, or as the one who will make Alex's life either happy or unhappy, is *me*. But the examples show that, without qualification, Coliva's criterion doesn't deliver the goods.

As advanced, the problem lies in the context-dependent relativity of knowing-wh questions, which Boër & Lycan (1986) pointed out. If, discussing the contemporary literary scene, I am asked 'Do you know who Rachel Cusk is?', I correctly answer 'Yes, I do.'; for I can mention the title of some of her books, I have read some of them, I can offer information about their contents and some appraisal. But if the question is put to me at a party at which she is present, I correctly answer instead 'No, I don't.'; for I don't know what she looks like; I cannot visually pick her out among the people at the party. Knowing-wh questions ask whether we can tender *contextually relevant* identifying information; what counts as relevant may change from context to context.

[8]Cf. also Coliva (2003: 428; 2006: 13, fn.; 2012: 24 fn.). Coliva's (and Echeverri's) account may have been inspired by Anscombe's remark, 'If you are a speaker who says "I", you do not find out what is saying "I"' (Anscombe 1975: 56). In the context, she seems to be taking this epistemic condition as just another way of putting the thought in the quotation at the start.

It doesn't matter for our purposes whether the sense of 'correctness' relevant to justify my description of the cases is just 'pragmatic', or instead 'semantic' (Braun 2006). In the first case envisaged in the previous paragraph, it is the ability to distinguish myself from other people in the picture by visual appearance that is contextually salient, and I lack it. In the second, it is identifying myself as either The Prince or The Beggar that is relevant, which I cannot do.

This is the first horn of the dilemma I anticipated in Section I, elaborating on the one that Sainsbury (2011: 253) articulates. In fact, Coliva suggests a way to qualify her knowing-wh criterion as it should be understood in her characterization of RG. The way I would put it (which may not be hers), what is at stake is not any substantive contextually salient identificatory knowledge, but identifying knowledge specifically required to fix *its semantic referent*: 'It is a rule for the competent use of "I" that one must use it to refer to *oneself*. But one cannot so much as be in a position to understand this rule unless one knows which person one is. This, in turn, implies that *qua* competent "I"-user, the subject knows which person is the semantic referent of "I"' (Coliva 2003: 428). She goes on to extend the characterization to thoughts: 'the first-person concept is the concept of *oneself* and one cannot have it unless one knows which person one is. Hence, there is a guarantee that *qua* possessor of the first-person concept, the subject knows which person that concept is a concept of' (*ibid*., 429).[9] I grant that this addresses the first horn objection; but it takes us to the second horn of our dilemma.[10]

Competent users of referential expressions (or deployers of a corresponding concept) should be able to identify their semantic values. This is, I submit, an adequate way of understanding Russell's 'Principle of Acquaintance' (García-Carpintero 2021; cf. Coliva 2017: 242). Alas, this doesn't distinguish self-reference by means of SELF$_1$ from self-reference by means of NN, against what Coliva needs. She says, 'if you believe "My hair is blowing in the wind", no matter how mistaken you are, and even if you are guilty of an error of misidentification, there is no possibility for you to go on sensibly to

[9]Coliva's (2017: 241–2) distinction between *individuation* and *identification* points in the same direction. Put once more in the way I understand these issues, the former amounts to *internally* identifying the referent of a concept, in ways constitutive of the concept; the latter, to identifying the referent of a concept in ways that are not constitutive of it, *externally*. See my related distinction between, respectively, *identification as a concept* and *identification as a premise* in my discussion of IEM (García-Carpintero 2018a: 3314).

[10]In sync with his austere proclivities—questionable for present concerns as I'll point out below, Section VI—Sainsbury (2011: 253) characterizes his own second horn of our dilemma thus: 'the requirement of identification or knowing-who places no constraints at all', which is closer to the way Coliva talks. On my view (see Section VI), this is inaccurate. On the second horn, the requirement does place a genuine constraint: to wit, that the term or concept in the relevant use is *internally* determined, picked out by constitutive reference-fixing information. The second horn raises for Coliva's proposal (and Echeverri's in Section III) the problem that, without further elaboration, the knowing-wh criterion thus understood doesn't distinguish SELF$_1$ from NN.

inquire which person has her hair blown by the wind' (Coliva 2017: 238–9). In contrast, however, she goes on, let's suppose that 'you correctly believe "Elena Cocò's hair is blowing in the wind"' (*ibid.*, 239) because you overheard a true assertion of it. Here 'you can sensibly wonder "Ok, Elena Cocò's hair is blowing in the wind, but which is this person (i.e. the one whose hair is blowing in the wind)?"' (*ibid.*) even if she *is* Elena Cocò.

There is an equivocation in this argument, which the brief discussion above on knowing-wh conditions helps us to spot. We can grant Coliva's last claim in the quote if we assume that what is being contextually required for identification of Elena Cocò is substantive information—knowing things about her such as how old she is, where she lives, what she does for a living, or what she looks like. But if (as we should, to have a non-equivocal comparison with the first-person case) we assume instead that we only need to have the information required to competently understand the name, it should be clear that the knowing-wh requirement is met and we cannot 'sensibly wonder' who Elena Cocò is. For Coliva's is a deferential use of the proper name 'Elena Cocò'; and it is quite enough to comply with Russell's Principle in that case that one intends to use the name deferentially, to refer to whoever those people were referring to by that name (García-Carpintero 2018b).[11] Paraphrasing the final part of a quotation from Coliva two paragraphs above, '*qua* competent "Elena Cocò"-user, the subject knows which person is the semantic referent of "Elena Cocò"'. The point extends to thoughts deploying 'Elena Cocò'.

To sum up: on the first horn of the dilemma, any contextually salient identifying information may be required for meeting knowing-wh conditions.

[11] Coliva (2017) elaborates on her view through the notion that first-personal concepts are 'stopping points of inquiry': 'once you entertain a first-person content, you have reached a stopping point of inquiry. If you are entertaining a genuinely first-personal content, you automatically know that it is you that have (or seem to have) the property in question and there is no room left for an inquiry concerning who has that property. Hence, to put it emphatically, what distinguishes first-person contents from impersonal ones is their 'luminosity' in one respect. Namely, they are luminous regarding the individuation of the subject who has (or seems to have) the property which gets ascribed to her in the proposition' (*ibid.*, 235). She goes on to grant that the same applies to indexicals and demonstratives, perhaps to numerals. But in fact, it extends to any competently deployed simple term/concept, including names, as the following example (which I owe to Daniel Morgan) shows. I think that there is this person, Ralph, whose presence in Leeds in a year's time is necessary and sufficient to avert a disaster. I might have been told this by an oracle and know nothing of Ralph's properties, not even that Ralph is not me. I might well say to someone who reassures me that I will be in Leeds in a year's time, 'Don't tell me that I will be in Leeds in a year's time. Tell me whether Ralph will be present in Leeds in a year's time'. Once I know that, my inquiry is over: I know the truth value of the claim that I care about. Names can thus also be 'stopping points of inquiry'. As a reviewer pointed out, Coliva's 'stopping points of inquiry' elaboration might suggest that, unlike me, she is after all not thinking of RG as distinguishing *de se* thought from others better than IEM. The dilemma I presented would then seem irrelevant. However, among other passages, the introductory paragraph in Coliva (2003: 416) clearly supports my interpretation. I understand that the 'stopping points of inquiry' elaboration only adds a necessary condition to *de se* thought and isn't meant to replace her *knowing-wh* account.

In this case, Coliva's knowing-wh characterization of RG fails because first-personal reference may not meet the condition. On the second, it is just *internal* identifying information constitutive of the relevant concept or meaning that is asked for. In that case, RG as characterized by Coliva doesn't single out first-personal reference because self-reference by means of names meets it as well.[12]

III. Echeverri's *Erotetic* characterization

Echeverri (2020: 478 fn.) mentions the problem that knowing-wh locutions are context-dependent, although he doesn't elaborate on why this is problematic for our goal of stating RG, as I just did. He offers an alternative characterization of RG, inspired nonetheless as he grants (*ibid.*) by Coliva's (cp. the 'Elena Cocò' quote above): '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 […]: 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' (*ibid.*, 477). Echeverri understands the 'impossibility' here in normative terms: raising the question while having the thought manifests *incompetence* or is *incoherent* given the nature of the concept (*ibid.*, 477–8).

Alas, although this revised epistemic criterion does better than Coliva's, a related dilemma also afflicts it. For the first horn, I'll consider a thought-insertion based example. In a discussion of such cases, Gregory (2016) notes that subjects may experience auditory verbal hallucinations not 'in' their own voice (cf. Wu 2012: 95–6). Gregory argues that these are not *inner speech* proper, but auditory imaginings. Now, let's suppose that a subject has an auditory experience of an assertoric utterance of *I should murder Lissi*.[13] Let's stipulate

[12]A reviewer objected that the dilemma is 'spurious … what is presented as a case in favor of the first horn conflates EM and RG. The subject may not know which person she is identical with (EM: "Am I this or that person?"—i.e. "I = this person?"), but she would still know who she is referring to by means of "I" even while asking herself whether she is the prince or the bagger.' This bluntly ignores the dialectics. I don't doubt that the speaker uttering 'I' in the examples in the first horn 'knows who she is'; this much I granted. The question is whether Coliva's knowing-who characterization of RG is correct. The reviewer is right that my examples in the first horn manifest vulnerability to misindentification; but they also show that, unqualified, Coliva's knowing-wh criterion doesn't pass muster. To show that, nonetheless, the speaker in them 'knows who she is', we may understand Coliva as I suggest in the second horn (which the reviewer doesn't object to); but now we fall prey to the objection I raise there.

[13]This is a real case; I don't know whether the account I'll offer fully fits it, but I don't need it. The report was 'One evening the thought *was given to me* electrically that I should murder Lissi' (Mullins and Spence 2003: 295; the authors take it from Jaspers' 1963 *General Psychopathology*). See Wu (2012) for nuanced distinctions between auditory verbal hallucination and inner speech in schizophrenia, consistent I think with my discussion below.

that this in fact is inner speech, appearing to the subject as made in a voice very much like her own.

In debates on whether thought insertion disproves the idea that the experience of one's own thoughts possesses IEM, it is pointed out that we must distinguish self-ascriptions of *authorship* of the relevant thoughts from self-ascriptions of *ownership* (cf., e.g., Coliva 2002: 31; Seeger 2015b; Duncan 2019: 305; Echeverri 2020: 495). Our subject is said to (coherently enough) disclaim the former but not the latter. Now, Gregory's point shows that this distinction doesn't suffice to aptly describe all cases. Our subject might well rationally wonder whether she is experiencing an assertoric utterance by herself in inner speech, or instead somehow overhearing or perhaps auditorily imagining someone else's in a voice very much like her own. Given that such imaginings are mental acts, the subject doesn't thus need to disclaim *producing* the judgement in some sense. The distinctions that we need are thus subtler. García-Carpintero (2024a) devises a threefold one between *committed authoring*—the occurrent thought manifests the subject's standing mental dispositions and she is open to endorse it given its rational character (whether judgement, intention, imagining, perceptual experience …); *producing*—being its relevant causal origin; and *entertaining*—being available to introspective self-ascription.[14]

As is well established—and the standard interpretation takes Perky's (1910) experiment to show, cf. Currie (2000: 180)—we sometimes find it difficult to distinguish (auditory) *perceptions* from *imaginings*—say, of bells tolling, or tunes that come doggedly to mind. Instead of wondering whether she is somehow overhearing an assertion by someone else, our subject might be querying whether she is imagining it; as Wu (2012: 88) reports, 'patients are not always sure that they are actually hearing the voices or whether they are only compelled to think them'. In that case, she may allow that she is its *producer*, given the well-established view that imaginings are typically things that we do—we can try to control them even if we fail in some cases of unbidden imaginings, as happens with compulsive physical actions (McGinn 2004: 13–4; cp. Wu 2012: 100–4). The subject may instead be doubting that she is the one *judging* (i.e., assertorically committing to the content), her *(committed) author*.

How does this affect our issue? If we assume a *thinker-reflexive* reference rule for 'I' as occurring in thought (Section VI), we still need to distinguish whether it is the thinker *as entertainer*, *producer* or *committed author* that determines the referent (Palmira 2020, 2022; Salje 2024). Now, as García-Carpintero (2024a: §3) points out, many times it is not a judgement, but a different type of thought that patients take to be 'inserted'; for instance, an order: *Kill Mom!* In fact, in the example we have been discussing, the judgement *I should kill Lissi* may express the patient's acceptance of a corresponding directive, *Kill Lissi!*

[14]Salje (2024) has a related if not entirely coincident threefold distinction; mine was inspired by an earlier version of this material.

Let's put it in the performative form, to highlight the most relevant instance of the first person for our purposes, *I am hereby ordering you to kill Lissi*. The way I interpret the case, if we take 'I' here to refer to the *author* and *primary producer*, she may coherently doubt that it refers to her—even if in fact the order is inner speech coming from herself and she is the producer and author.[15] Thus understood, this is a case in which a subject is experiencing what in fact is a use (token) of her indexical self-concept (an instance of 'I' in her inner speech) while it is rational for her to ask herself at the same time, *Does 'I' refer to me?*[16] This corresponds to the first horn of the dilemma above for Coliva's articulation of RG, now addressed to Echeverri's related erotetic account.[17]

Now, Echeverri might plausibly contend that I am being too liberal in assuming that our subject is truly deploying a token of SELF_I.[18] Perhaps we can think of mental reference the way Searle (1969) suggests for the linguistic case, as an 'ancillary' speech act through which one means a specific singular proposition. We can then argue that, to the extent that the subject leaves open that she might be perceiving or imagining an utterance by someone else,

[15]Considering instead of 'I should kill Lissi' the command that this assumes helps because, if the declarative is taken to express its acceptance by the thinker, 'I' in it refers to her in the three roles, as *entertainer*, *producer* and *committed author*. 'Imperative thoughts do not feature any token of the I-concept', Echeverri (2020: 497) says. But the performative equivalence reveals that the first person does tacitly occur in imperatives, in the role of the agent of the expressed directive. The performative version explicitly features the relevant 'I' so that we can address Echeverri's articulation of RG; but what the subject is ultimately wondering is whether she is self-addressing the order *Kill Lissi!* to herself, or it is somebody else who is giving it—which may be weird, but still is a coherent question to raise, I think.

[16]Discussing the issue, Peacocke (2008) says: 'Consider the schizophrenic subject who suffers the experience labelled "thought-insertion", and to whom it seems that in having the thoughts occurring to him, he is overhearing someone else's thoughts. The thoughts in respect of which he has such a disturbing consciousness can be first-person thoughts. But this thinker does not even believe, let alone know, that the first person in such thoughts refers to him' (*ibid.*, 89). This is not true of 'I should kill Lissi' if 'I' refers to the entertainer, and it is not true either if it refers under the other two roles if the utterance expresses acceptance of the directive. He also says, 'the schizophrenic subject lacks […] action-awareness of the thoughts that occur to him' (Peacocke 2008: 276), which for the same reason doesn't need to be the case either. Nonetheless, I take it that the case Peacocke wants to suggest (and Echeverri engages with, see next fn.) is very close to the one I have developed. Salje (2024) has similar memory illustrations.

[17]On the envisaged scenario, the thinker leaves open the possibility that she is the one tokening 'I', and hence self-referring as Echeverri (2020: 486) defines this, as 'an act of producing a token of the I-concept'. But she is not committed to the request, nor referring to herself in indicating the content of that commitment; this is how the possibility remains open that the thinker 'can (coherently) take the referent of her own token of I to be different from herself', cp. Echeverri (2020: 480). Echeverri's 'intuition' is to deny that the case is coherent and possible (*ibid.*, 496–7), but hopefully my presentation makes it intuitive enough, putting pressure on the denial of its coherence. Echeverri's alternative suggestion is that, to the extent that the case is coherent, his erotetic RG is not violated (*ibid.*, 497–8). This is, I take it, the move that the second horn of the dilemma below targets. Palmira (2022) also questions Echeverri's account by means of thought-insertion cases, albeit with different considerations.

[18]It is unclear to me how Echeverri would characterize these cases, in part because he just assumes the standard twofold distinction, cf. his discussion of Peacocke's discussion and his Mary cases (Echeverri 2020: 499–500, 498). Thanks to Michele Palmira here.

she is not truly referring with 'I' nor thereby self-referring. The worry now is how 'deploying a true token of SELF$_I$' is to be explained, which takes us to the second horn of the dilemma. It may well be that, under any appropriate understanding of what it takes for a referential concept to be truly 'used', the erotetic condition also fails in uses of NN. Let 'EC' be a proper-name-like concept for the thinker in the previous case, and let's suppose that the situation now concerns what in fact is an inner speech utterance of *EC should murder Lissi*. Assuming stricter use conditions for the deployed concept, why should it be coherent for her to wonder at the same time, *Does 'EC' refer to EC/that person?* Why would she still be competently deploying that concept?[19]

Drawing on Anscombe's suggestions (fn. 8), Coliva and Echeverri define RG by means of epistemic conditions on concept-possession. This might seem apposite, because the conditions through which they articulate their criteria are pre-theoretical enough to meet our desiderata for an account of RG. Our critical discussion shows however that their proposals don't work when it comes to articulating the truly distinctive feature of indexical self-reference that RG points towards. I'll pursue instead the 'Cartesian' tack that Anscombe herself suggests.

IV. The real Real Guarantee

In a recent paper, Salje (2020: §2) offers a characterization of the intuition expressed in the text by Anscombe quoted at the start that gives us what we need—an account of the pre-theoretical claim RG that fares better than Coliva's and Echeverri's and can thus deliver the explanatory rewards I set up for it in Section I. She calls it 'Special Insight':

> *Special Insight*: In virtue of being the thinker of a conscious I-thought, a subject has privileged non-inferential epistemic grounds for first-personal knowledge that the referent of their thought exists. (Salje 2020: 739)

A 'conscious I-thought' is the sort of fundamentally first-personal attitude we are taking RG to be distinctive of—one deploying SELF$_I$, expressed by

[19]The reviewer mentioned above, fn. 12, objects that '[t]he subject may not know who EC (who should murder Lissi) is, while deploying the concept in thought. If, in contrast, she were thinking "I should murder Lissi", she could not fail to know which person she is referring to by means of "I"—i.e. herself. This shows that the deployment of the first-person concept guarantees the fulfillment of RG, while the deployment of other singular concepts, which may also have the same reference as "I", does not'. Once more, this egregiously ignores the dialectics. I agree that a good account of RG will show that the subject 'could not fail to know which person she is referring to by means of "I"'; but the issue at stake here is whether Echeverri's erotetic account manages to do that. If one grants that in my thought-insertion-like example the subject is in fact referring to herself with 'I', then the erotetic criterion fails, because the subject *can* rationally raise the question (first horn). If one contends that, under the circumstances, she is not, then, under the same circumstances, she wouldn't be with 'NN' either (second horn).

English speakers with first-personal pronouns. Salje's characterization is not *pre-theoretical* in a sense: it deploys notions for which she assumes philosophical accounts, including 'first-personal knowledge' (knowledge whose 'referent is thought of in a first personal way' (*ibid*) and 'privileged' ('a maximally weak version of a familiar privileged access, or epistemic asymmetry thesis. It is weak because it makes no claims to the self-intimating nature, infallibility, incorrigibility, or even the superiority of the thinker's side of this asymmetry and is largely silent on how best to understand the introspective mechanisms underpinning the epistemic grounds' (*ibid*). There are also 'non-inferential', 'epistemic grounds', 'in virtue of', which she doesn't gloss but understands as explicated in current epistemological debates. But this applies as much to Coliva's and Echeverri's accounts, or for that matter to any other decent philosophical account of a pre-theoretical notion, so it doesn't disqualify it.[20] The presupposition of a referent for conscious I-thoughts is of course not neutral; it is incompatible with views that favour Lichtenbergian, subject-free descriptions of the knowledge subjects are supposed to have privileged access to, like the one Anscombe herself appears to favour. By relying on previous discussions, Salje (*ibid.*, 740–1) makes a very good case that we should reject such views.[21]

Salje takes the view that *Special Insight* is (a good characterization of) RG as an objection. But first, as a formulation of RG *Special Insight* improves on extant accounts. Crucially, Salje convincingly shows (*ibid.*, 742–3) that SELF_I differs from NN vis-à-vis *Special Insight*—the crucial explanatory virtue of RG that, I have been arguing, Coliva's and Echeverri's accounts fail to secure. As an account of RG, *Special Insight* prevents our dilemma-like objection because, as Salje shows, referring to ourselves by means of a proper-name-like concept lacks the privilege that *Special Insight* articulates. Secondly, Salje's reason for distinguishing *Special Insight* from RG is that the latter is a *semantic* feature of *de se* thought, while the former is *epistemic*. But this presupposes a non-existent dichotomy. She grants that '*Special Insight* is arguably the epistemic face of guaranteed reference: while the guaranteed reference thesis says that I-thoughts cannot fail of reference, Special Insight says that the I-thinker always has epistemically distinctive grounds to know that her conscious I-thought successfully refers' (*ibid.*, 742). But it is in such epistemic terms that I (as well as Coliva and Echeverri) have been understanding the phenomenon in need

[20]For a related case, the lying vs. misleading distinction is pre-theoretical, and as such it has been taken as a datum to explore philosophical accounts of the semantics vs. pragmatics distinction (Saul 2012; Michaelson 2016; cf. García-Carpintero 2023 for discussion). But it is controversial how to philosophically characterize it and which notions to use in the task.

[21]To uphold RG also requires dealing with Evans's (1982: 249–5) alleged cases of failure of reference with 'I'. I cannot go into how the proposals to account for RG presented below—mine in particular—would dispose of his arguments; suffice it to say that Evans is assuming a very problematic self-location requirement for successful reference (O'Brien 2007: 37–8).

of elucidation; Anscombe clearly assumes this in the initial quotation. Quite in general, good semantic proposals should be supported by a grounding metasemantics, and any adequate metasemantics will mention epistemic facts.

Coliva's and Echeverri's accounts offer recipes to establish that a thought involves genuine self-reference; but, as the previous two sections have shown, the recipes don't faithfully fulfil their goals. The account based on Salje's notion offers a more trustful one: 'Can one be certain that the referential notion one deploys has a referent?' This criterion can also be argued to be pre-theoretical: given folk reactions to the *cogito*, ordinary people would find it intuitively applicable. Different philosophical accounts will explain it in their own ideology or explain away its pull otherwise. The reader might worry that characterizing RG as I have suggested would confirm Anscombe's point that only a Cartesian ontology can validate RG if 'I' refers. But this is not so. I'll show this in the final two sections by referring the reader to a range of views that can account for the datum in their proprietary terms, including one that I have been advocating (García-Carpintero 2002, 2015, 2016, 2017, 2018a).

V. Grounding the *Real Guarantee*: inner awareness

An account of *de se* thoughts that I have been defending accounts for RG without committing to a Cartesian ontology—an ontology that posits a referent of 'I' that neither is physical nor has a constitutive link to anything physical—as Anscombe feared. The view is a token-reflexive account of indexical reference that relies on a non-intentional acquaintance relation with one's conscious states; I'll motivate the latter in this section, and I'll outline the former in the next. Other accounts can explain RG, so this is not an argument for the view, which would require wider abductive considerations in any case. It is only intended to show that, as characterized, RG can be accounted for, and that it can be accounted for without incurring Cartesian commitments.

Phenomenal consciousness constitutively features *qualia*—properties of experiences such that *there is something it is like* for a subject in virtue of having them (Nagel 1974). *Qualia* are thus discriminating specific features of conscious states. Philosophers since Aristotle have also discussed a general feature common to all paradigmatic phenomenally conscious states:

> There is something it is like to taste chocolate, and this is different from what it is like to remember what it is like to taste chocolate, or to smell vanilla, to run, to stand still, to feel envious, nervous, depressed or happy, or to entertain an abstract belief. All of these different experiences are, however, also characterized by their distinct first-personal character. The what-it-is-likeness of phenomenal episodes is properly speaking a what-it-is-like-*for-me*-ness. This for-me-ness doesn't refer to a specific experiential quality like sour or soft, rather it refers to the distinct first-personal givenness of experience. It

> refers to the fact that the experiences I am living through are given differently (but not necessarily better) to me than to anybody else. (Gallagher and Zahavi 2021: §1)

This characterization leaves the nature of the general feature quite open. Different terms have been used for it, including Block's (1995) 'me-ishness', Kriegel's (2009) 'inner awareness', and Gallagher & Zahavi's 'for-me-ness'.[22] Borrowing from the phenomenological school, Zahavi (2005) and Boner et al. (2019) use the descriptively accurate 'pre-reflective self-awareness'. Here I'll mostly use Levine's (2001: 6–7) also evocative but less unwieldy terminology, contrasting 'qualitative character' and '*qualia*' with 'subjective character' and 'subjectivity'.

We should distinguish ('pre-reflective') *subjective character* from *introspection*, by which I understand a (reflective) conscious judgement about features of one's mental life. Let's consider a perceptual state: visually experiencing a pink cube in front of one. We could introspect the perceptual state and its features, its perceptual mode (thus discriminating it from, say, a visual imagining with the same content), the pinkishness in it, whether the length of the cube edges appears to be shorter than the distance at which the cube appears to be, and so on. The introspective stance is another conscious state with its own *quale*, requiring conceptual capacities not needed to have the experience, and directing attention to it. The subjective character of the perceptual state is thought to be a feature that it has whether or not it becomes the target of introspective reflection. It is hence a feature that, by being conscious, introspective states also have, even when they are not the target of a further, second-order introspection.

Different philosophical accounts of subjectivity have been offered. On a deflationary view subjectivity is a feature of phenomenal experiences that, by itself, doesn't make a distinctive additional contribution to phenomenal character—to what it is like for its subject to have that conscious experience. It is only through introspection that we gain a conscious awareness of it (Stoljar 2018). Hume's famous incapacity to find himself in experience is a phenomenological departing point for the deflationary line. It is appealing to a naturalistic stance on phenomenal consciousness that emphasizes the 'transparency' of conscious experience and understands it in representational terms, taking *qualia* to be properties of represented items. Stoljar's (2021: §15) main consideration for the deflationary view mentions this alleged datum of transparency. Now, these accounts accept a view for which Block (2015) offers compelling considerations: that there are unconscious states with representational features, perhaps the very same as the ones had by conscious

[22]There are more, cf. Guillot (2017: 25), Byrne (2004: §3). Farrell & McClelland (2017), Boner et al. (2019) and García-Carpintero & Guillot (2023) are recent compilations on this topic. For my (biased) presentation I am borrowing here from García-Carpintero & Guillot's introduction, for which I thank my co-author.

states.[23] What makes the difference? Relatedly, what distinguishes my own conscious states from those I ascribe to others? The quotation above from Gallagher & Zahavi (2021) suggests a consideration for a more robust view of subjectivity based on that observation.

To fill up such perceived explanatory lacunae in deflationism, robust theories of subjectivity elaborate on an intuition nicely captured by H. H. Price in this famous passage:

> When I see a tomato there is much that I can doubt. I can doubt whether it is a tomato that I am seeing, and not a cleverly painted piece of wax. I can doubt whether there is any material thing there at all. Perhaps what I took for a tomato was really a reflection; perhaps I am even the victim of some hallucination. One thing however I cannot doubt: that there exists a red patch of a round and somewhat bulgy shape, standing out from a background of other colour-patches, and having a certain visual depth, and that this whole field of colour is directly present to my consciousness. What the red patch is, whether a substance, or a state of a substance, or an event, whether it is physical or psychical or neither, are questions that we may doubt about. But that something is red and round then and there I cannot doubt. (Price 1932: 3)

Price is, I take it, making intuitively salient the 'real presence' of the self in phenomenally conscious states that our initial quotations from Anscombe also point out and RG captures on the proposal in Section IV. The quotation highlights what to me is the main reason for questioning the alleged datum of transparency—the primary intuitive phenomenological datum for subjectivity. While the painful condition of my tooth that we may take my toothache to represent (and the tooth itself) may fail to be there alongside my experience, the pain that I feel cannot fail to exist; it is there, present, part of the actual world as much as the feeling itself. RG targets this as a datum in need of explanation or justified dismissal.[24]

Two robust, non-deflationary views explain the data, making the connection with RG clear. On representationalist views (Kriegel 2009; Zahavi 2018; Coleman 2019) subjectivity consists in the fact that phenomenally conscious states represent themselves and their qualitative features, perhaps also the

[23] Cf. also Quilty-Dunn (2019).

[24] Stoljar's (2021: §13) 'argument 9' is based on this point. He enlists quotations from Chalmers's work, in which Chalmers makes the point in terms familiar from Peacocke's (1983) distinction between *sensational* and *representational* properties of experiences, and Block's (2003) arguments for 'mental paint'—say, blurry visual experiences, or attending to the variable features of visual experiences when the represented features (size, colour, and so on) on which we normally focus our attention are kept identical by perceptual constancy mechanisms. I take Price's quotation to zoom in on the crucial datum, close to RG. García-Carpintero (2002) uses these points in an argument for *sense data* taken as I think we should understand them, as theoretical entities in ontology, cf. also Lowe (1986, 2008); Lowe (2008: 69) reproduces Price's point. Note that Price's sense data are 'red and round' only in a metonymical extended sense; with Peacocke (1983), properly speaking we should say that they are *red*' and *round*'.

subject. On the other substantive account of subjectivity, it is explained as a real (as opposed to intentional) relation of *acquaintance* relating the subject to those very items (Williford 2015; Raleigh 2019; Duncan 2018, 2019, 2021). Subjective character is 'a kind of implicit acquaintance with oneself, or background self-familiarity' (Schear 2009: 96). The two robust views allow that the features of phenomenally conscious states that the states self-represent or which their subjects are acquainted with become contents of conscious introspective states that target them (Chalmers 2010; Gertler 2012).

A main motivation for deflationary views comes from well-motivated worries that robust views on phenomenal consciousness might be anti-naturalist; Pelczar (2019) defends a recent version, on which objects represented in perceptual experiences are ungrounded dispositions to produce conscious experiences. Such views presuppose the representational view of subjectivity because states representing 'external' objects are not just ontologically but also epistemically grounded on states presenting inner features (Farkas 2013). In addition to ontological worries, they thus also raise 'veil of perception' epistemic concerns. But as Lowe (1986, 2008) points out, the acquaintance view doesn't require any of this. It claims that, ontologically, awareness of external objects is mediated by awareness of internal features of our conscious experiences. This needs not be *epistemic* mediation. *Qualia* proper (the features of conscious states we are acquainted with in inner awareness) might be seen as theoretical entities (Sellars 1963), whose precise character (whether events or particulars, types or tokens), as Price intimates, we learn about from the explanatory roles they play. A view along these lines allows for the limitations in our knowledge of *qualia* that Williamson's (1996) anti-luminosity considerations highlight, and blocks Wittgenstein's Private Language argument (García-Carpintero 2002, 2003).

Richard Wollheim's (1998) influential account of depiction theorizes an experience of *seeing in* endowed with a dual character, *twofoldness*. The way I understand it (cf. Stecker 2013: 148–9; Terrone 2020: 180) this is a unique experience occurring when we are aware of experiencing a picture—a painting or an image in a screen—that allows its subjects to attend to two different constitutive features thereof: a meaning-vehicle, a co-present two-dimensional image in egocentric space; and an imagined three-dimensional represented situation.[25] The acquaintance view ascribes a similar *twofoldness* to all conscious experiences. They constitutively involve acquaintance with some of its currently instantiated features, and a represented intentional content. In closing, I'll show how this affords a straightforward account of RG.

[25]Bengson et al. (2011) offer a similar 'Dual Character' account of perceptual experience.

VI. Accounting for the *Real Guarantee*

The token-reflexive account of *de se* thought that I favour can be usefully contrasted with the one Sainsbury (2011) articulates here for the linguistic case:

> there's no more to understanding a token of 'I' […] than being able to apply to the token the rule: English speakers should use 'I' to refer to themselves as themselves […] Some identificatory work is presupposed, but that is identification of an utterance and not, in any substantive sense, identification of an utterer. Since token utterances have their utterers essentially, in a sense we have identified the utterer once we have identified the utterance; this exemplifies the 'non-substantive' or trivial notion of identification of the utterer: he or she is identified simply as the utterer. (Sainsbury 2011: 254–5)

Sainsbury considers Anscombe's (1975: 47–8) concern that the qualification 'as themselves' in his statement of the self-reflexive rule (crucially required to distinguish the targeted indexical 'mode of meaning' oneself) deprives the proposal of explanatory power. He answers it by providing an allegedly explanatory elaboration: he qualifies *referring to oneself* as 'an application of the rule: use 'I' to refer to yourself', which 'precludes referring to oneself in [an] ignorant way' (*ibid.*, 257). This is ok as far as it goes; but it clearly doesn't go far enough. First, some account must be given of what it is for a subject to follow a rule in a way that prevents the envisaged ignorance. It is not enough to regularly act in accordance with the rule; the rule must somehow 'guide' such acts. Secondly, while Sainsbury's account is offered for linguistic acts, what we are after is one for *de se* thoughts. How can we extend it to the mental realm? In the linguistic case the rule might have a social character; but what about the mental case? Who or what enforces it? What does *being guided* by it amount to? Sainsbury doesn't say; but he intends his account to apply to thoughts, as his attempt to explain IEM mentioned below shows.

In my work (García-Carpintero 2000, 2015, 2016, 2017, 2018a), I have answered these questions in a straightforward way. Assuming that there is cognitive phenomenology, whether reducible to experiential phenomenology or not, I extend Peacocke's (1983) distinction between sensation and representation to thoughts. I assume the twofoldness of conscious states outlined in Section V. They feature an intentional awareness of contents, and a non-intentional awareness of, or acquaintance with, features of content-vehicles, depictive ones in the case of visual or auditory experiences, linguistic in the case of cognitive phenomenal states—inner speech offers a central illustration. In this way, we may be non-intentionally aware of a token of the SELF$_1$ concept/term, which provides grounds for extending to thoughts the sort of account that Sainsbury offers.[26] Needless to say, none of this is

[26]Because of its twofoldness—its reliance on self-acquaintance—my appeal to the reflexive rule is not 'bare' (Palmira 2020, 2022), but rather presumes a very specific 'gloss', as he puts it.

philosophically unproblematic; a relevant motivation for it is that it affords a clear-cut account of RG.[27]

Sainsbury's view aims to be deflationary, but I think we should question his claims in this regard: 'The notion of an "essentially indexical thought" might wrongly be taken to refer to a specific kind of content […] No content is distinctive of self-knowledge' (*ibid.*, 255–6). Against this, I have argued that we should cash out the token-reflexive material fixing the referents of indexicals as a linguistically triggered presupposition, thereby not part of 'at issue' content. I thus agree to some extent with what Sainsbury says in the quoted passage. But in the pre-theoretical sense that the 'content' metaphor gestures at, the background token-reflexive presuppositions are still part of the content of the full representational act. I extend the view to the mental, with background beliefs playing the role that the *common ground* performs for linguistic presuppositions, which affords a Perryan view on the *de se*.[28] I argue that the role that non-intentional awareness of token self-concepts plays in them makes such states private and non-shareable in a less deflationary sense than Sainsbury allows.

It may well be that self-reflexive accounts can be developed (as in any case I have shown Sainbury's should be) in less controversial directions. Thus, Bermúdez (2016), Howell (2006), Longworth (2013), and Verdejo (2018) offer accounts meant to preclude incommunicability. Echeverri's (2020, 2021) proposal is another example. He provides a functionalist account of the differences between SELF_I and NN. He points out that, if an *acquaintance* account is just one that appeals to a self-reflexive rule in a non-descriptivist way (i.e. without making SELF_I synonymous with the self-reflexive condition), the one by means of which he explains RG as he understands it (Echeverri 2020: §3) counts as such, without implying non-shareability of the SELF_I concept. Now, to achieve the explanatory virtues of positing acquaintance in the robust sense outlined in Section V (in particular, to circumvent circularity objections to self-reflexive rule accounts like the one by Anscombe (1975: 47–8) that

I take our views to be very similar. García-Carpintero (2016: 192; 2018a: 3320, 3324) questions Peacocke's efforts to make do, like Sainsbury, with a less committal reliance on the reflexive rule to account for self-reference and self-awareness.

[27] In addition to the positive *Special Insight*, Salje (2020) posits a negative epistemic feature of I-thoughts, *Ordinary Ignorance*: 'A subject does not, in virtue of being the thinker of a conscious I-thought, have privileged noninferential grounds for knowledge about the nontrivial properties of the referent of their thought' (*ibid.*, 743). Deflationists on the significance of the self-reflexive rule like Sainsbury might agree with this; but my own view at the very least would qualify it. It is true that we need very little knowledge of ourselves for self-reference; but the acquaintance on which I rely needs to get hold of some non-trivial qualitative properties, available for introspective knowledge. Thanks to Michele Palmira for suggesting adding this observation. I still agree with the main point that Salje makes in her paper (see fn. 2 above). She claims that the combination of the two features of I-thoughts generates a cognitive illusion 'in which the […] self […] appears to be by its very nature substantively unknowable' (Salje (2020: 739), 'a mysterious or otherworldly object' (*ibid.*, 747). The appeal to acquaintance to explain *Special Insight* doesn't undermine this view, which I find compelling.

[28] Cf. García-Carpintero (2016: 181–2), Torre (2018: 183–4).

Sainsbury addresses), Echeverri contends that deployment of a token of SELF_I in thought 'is underwritten by a basic, self-referring act' (*ibid.*, 486), where *basic self-referring act* is a primitive notion. He worries that promoters of robust acquaintance may complain that this 'comes very close to an acquaintance view' (*ibid.*, 487). My worry is rather that this notion shouldn't be taken as primitive but explained as my proposal allows.[29] But a proper discussion of these issues should be left for another occasion.

If developed as suggested, Sainsbury's account of *de se* thoughts explains RG. The thinker of a conscious I-thought has privileged non-inferential epistemic grounds for first-personal knowledge that the referent of their thought exists because reference is fixed by acquaintance with a real, constitutive part of the subject herself. Paraphrasing what Sainsbury says in the quotation above, assuming (mental) token utterances have their subjects essentially, the acquaintance with them that fixes reference to the subject thereby offers non-inferential grounds for knowledge that the referent exists. This might be seen as the basis for *cogito*-like inferences (Peacocke 2012, 2021). Sainsbury (2011: §6) also shows how reliance on the self-reference rule explains IEM: 'I can't use the first person pronoun (in the normal way) yet erroneously refer to someone or something other than myself'. Once more, however, I would argue that for the explanation to work the substantive elaboration I have outlined is needed.[30]

In non-deflationary ways, Palmira's, Guillot's (2023) and my account are also prima facie plausible. To the extent that these proposals rely (as Sainsbury puts it in the above quotation) on the identification of token 'utterances', they all confirm Anscombe's suspicion that 'this reference could only be sure-fire if the referent of "I" were both freshly defined with each use of "I", and also remained in view so long as something was being taken to be *I*' (Anscombe 1975: 57). In addition to Anscombe's point that the referent of the token SELF_I concept is 'freshly defined with each use', Palmira's, Guillot's and my acquaintance-based view also substantiate the related point she makes in the quotation at the start that it 'guarantees the existence because it guarantees the presence, which is presence to consciousness'.

[29]I am similarly sceptic that, to understand Anscombe's no-reference claim, Doyle (2016), Haddock (2019) and Stainton (2019) have truly identified a position halfway between the standard interpretation that assimilates 'I' to expletive 'it', and robust views like Evans's that simply reject the claim. These writers assume a deflationary 'no-reference' understanding of the reflexive rule (on which they rely to explain how 'I' works according to Anscombe in their interpretation), which I don't think is available. Wiseman (2017) provides an alternative account, connecting Anscombe's views on 'I' to her work on intention. But even granting that her suggestions may work for self-ascriptions of intentions, I cannot see how they extend to, say, 'I have toothache', which Anscombe clearly meant to cover.

[30]García-Carpintero (2018a, 2024a, 2024b) offers a semantic account of IEM along these lines, arguing that it captures the distinction between non-IEM, merely *de facto* and *de jure* IEM *de se* thoughts (cp. Coliva 2017: 240, 244 for scepticism about such accounts).

The availability of these proposals further shows that the account of RG I have offered improves on Coliva's and Echeverri's, by elucidating the difference between SELF$_I$ and NN. Unlike Heimson's 'freshly defined with each use' SELF$_I$ concept, his 'David Hume' NN concept clearly fails to satisfy RG: it just fails to refer to anything currently in existence. Even though speaking *sub specie aeternitatis* it does refer to Hume, we could easily concoct cases of subjects with 'Vulcan'-like NN concepts. I conclude that we should stick to *Special Insight* to state the RG condition that good accounts of *de se* thoughts should either validate or explain away, thereby answering recent skepticism about their intuitively distinctive character.[31]

References

Anscombe, G. E. M. (1975) 'The First Person', in S. D. Guttenplan (ed.) *Mind and Language*, pp. 45–65. Oxford: Clarendon Press.

Bengson, J., Grube, E., and Korman, D. (2011) 'A New Framework for Conceptualism', *Noûs*, 45: 167–89.

Bermúdez, J. L. (2016) *Understanding 'I'. Language and Thought*, pp. 199–220. Oxford: OUP.

Block, N. (1995) 'On a Confusion about a Function of Consciousness', *Behavioral and Brain Sciences*, 18: 227–47.

Block, N. (2003) 'Mental Paint', in M. Hahn and B. Ramberg (eds) *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, pp. 165–200. Cambridge, MA: MIT Press.

Block, N. (2015) 'The Anna Karenina Principle and Skepticism about Unconscious Perception', *Philosophy and Phenomenological Research*, 93: 452–9.

Boër, S. E. and Lycan, W. (1986) *Knowing Who*. Cambridge, MA: MIT Press.

Boner, M., Frank, M., and Williford, K. (eds.) (2019) 'Senses of Self: Approaches to Pre-Reflective Self-Awareness', *ProtoSociology*, 36.

Braun, D. (2006) 'Now You Know Who Hong Oak Yun Is', *Philosophical Issues*, 16: 24–42.

Byrne, A. (2004) 'What Phenomenal Consciousness Is ', in R. Gennaro (ed.) *Higher-Order Theories of Consciousness: An Anthology*, pp. 203–25. Amsterdam: John Benjamins.

Campbell, J. (1999) 'Immunity to Error through Misidentification and the Meaning of a Referring Term', *Philosophical Topics*, 26: 89–104.

Cappelen, H. and Dever, J. (2013) *The Inessential Indexical*. Oxford: OUP.

Castañeda, H. (1966) ''He': A Study in the Logic of Self-Consciousness', *Ratio*, 8: 130–57.

Castañeda, H. (1968) 'On the Logic of Attributions of Self-Knowledge to Others', *Journal of Philosophy*, 65: 439–56.

Chalmers, D. (2010) *The Character of Consciousness*. Oxford: OUP.

Coleman, S. (2019) 'Natural Acquaintance', in J. Knowled and T. Raleigh (eds.) *Acquaintance*, pp. 49–74. Oxford: OUP.

Coliva, A. (2002) 'Thought Insertion and Immunity to Error through Misidentification', *Philosophy, Psychiatry and Philosophy*, 9: 27–34.

[31] Financial support for my work was provided by MICIU/AEI/10.13039/501100011033, research projects [PID2020-119588GB-I00, CEX2021-001169-M], and through the award 'ICREA Academia' for excellence in research, 2018, funded by the Generalitat de Catalunya. The paper was presented at the Ligerz 'Apperance and Reality' workshop, February 2022, and received helpful comments there. Thanks to Annalisa Coliva, Adrian Haddock, Max Kölbel, Rory Madden, Mike Martin, Daniel Morgan, Michele Palmira, Léa Salje, Carlota Serrahima, and Víctor Verdejo for very useful comments and suggestions, and to Michael Maudsley for the grammatical revision.

Coliva, A. (2003) 'The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee', *Journal of Philosophy*, 100: 416–31.
Coliva, A. (2006) 'Error through Misidentification: Some Varieties', *Journal of Philosophy*, 103: 403–25.
Coliva, A. (2012) 'Which 'Key to all Mythologies' about the Self? A Note on Where the Illusions of Transcendence Come from and How to Resist Them', in S. Prosser and F. Recanati (eds) *Immunity to Error through Misidentification: New Essays*, pp. 22–45. Cambridge: CUP.
Coliva, A. (2017) 'Stopping Points: "I", Immunity and the Real Guarantee', *Inquiry*, 60: 233–52.
Currie, G. (2000) 'Imagination, Delusion and Hallucinations', *Mind and Language*, 15: 168–83.
Doyle, J. (2016) '"Spurious Egocentricity" and the First Person', *Synthese*, 193: 3579–89.
Duncan, M. (2018) 'Subjectivity as Self-Acquaintance', *Journal of Consciousness Studies*, 25: 88–111.
Duncan, M. (2019) 'The Self Shows up in Experience', *Review of Philosophy and Psychology*, 10: 299–318.
Duncan, M. (2021) 'Acquaintance', *Philosophy Compass*, 16: e12727.
Echeverri, S. (2020) 'Guarantee and Reflexivity', *Journal of Philosophy*, 117: 473–500.
Echeverri, S. (2021) 'Putting I-thoughts to Work', *Journal of Philosophy*, 118: 345–72.
Evans, G. (1982) *The Varieties of Reference*. Oxford: Clarendon Press.
Farkas, K. (2013) 'Constructing a World for the Senses', in U. Kriegel (ed.) *Phenomenal Intentionality*, pp. 99–115. Oxford: OUP.
Farrell, J and McClelland, T. (2017) 'Editorial: Consciousness and Inner Awareness', *Review of Philosophy and Psychology*, 8: 1–22.
Gallagher, S. and Zahavi, D. (2021) 'Phenomenological Approaches to Self-Consciousness', *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/spr2021/entries/self-consciousness-phenomenological>, accessed 28 March 2022.
García-Carpintero, M. (2000) 'A Presuppositional Account of Reference-Fixing', *Journal of Philosophy*, XCVII: 109–47.
García-Carpintero, M. (2002) 'Sense-data: the Sensible Approach', *Grazer Philosophische Studien*, 62: 17–63.
García-Carpintero, M. (2003) 'Qualia that it Is Right to Quine', *Philosophy and Phenomenological Research*, 67: 357–77.
García-Carpintero, M. (2015) 'De Se Thought', in S. Golberg (ed.) *Oxford Handbooks Online*. Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780199935314.013.61, Retrieved 12 November 2015.
García-Carpintero, M. (2016) 'Token-reflexive Presuppositions and the De Se', in M. García-Carpintero and S. Torre (eds) *About Oneself*, pp. 179–99. Oxford: OUP.
García-Carpintero, M. (2017) 'The Philosophical Significance of the *De Se*', *Inquiry*, 60: 253–76.
García-Carpintero, M. (2018a) '*De Se* Thoughts and Immunity to Error through Misidentification', *Synthese*, 195: 3311–33.
García-Carpintero, M. (2018b) 'The Mill-Frege Theory of Proper Names', *Mind*, 127: 1107–68. CrossRef]
García-Carpintero, M. (2021) 'Reference-fixing and Presuppositions', in S. Biggs and H. Geirsson (eds.) *Routledge Handbook of Linguistic Reference*, pp. 179–98. London: Routledge.
García-Carpintero, M. (2023) 'Lying vs. Misleading, with Language and Pictures: The Adverbial Account', *Linguistics and Philosophy*, 46: 509–32.
García-Carpintero, M. (2024a) 'Is Conscious Thought Immune to Error through Misidentification?', *Philosophical Psychology*, https://doi.org/10.1080/09515089.2024.2351535
García-Carpintero, M. (2024b) 'Memory-based Reference and Immunity to Error through Misidentification', *Synthese*, 204: 1–14. https://doi.org/10.1007/s11229-024-04664-2
García-Carpintero, M. and Guillot, M. (2023) 'Introduction: Views about Self-Experience', in M. García-Carpintero and M. Guillot (eds.) *Self-Experience: Essays on Inner Awareness*, pp. 1–23. Oxford: OUP.
Gertler, B. (2012) 'Renewed Acquaintance', in D. Smithies and D. Stoljar (eds.) *Introspection and Consciousness*, pp. 93–127. Oxford: OUP.
Gregory, D. (2016) 'Inner Speech, Imagined Speech, and Auditory Verbal Hallucinations', *Review of Philosophy and Psychology*, 7: 653–73.

Guillot, M. (2017) 'I Me Mine: on a Confusion Concerning the Subjective Character of Experience', *Review of Philosophy and Psychology*, 8: 22–53.
Guillot, M. (2023) 'The Phenomenal Concept of Self and First-person epistemology', in M. García-Carpintero and M. Guillot (eds) *Self-Experience: Essays on Inner Awareness*, pp. 223–49. Oxford: OUP.
Haddock, A. (2019) ''I am NN': a Reconstruction of Anscombe's 'The First Person'', *European Journal of Philosophy*, 27: 957–70.
Howell, R. J. (2006) 'Self-Knowledge and Self-Reference', *Philosophy and Phenomenological Research*, 72: 44–69.
Kriegel, U. (2009) *Subjective Consciousness: A Self-Representational Theory*. Oxford: OUP.
Levine, J. (2001) *Purple Haze*. Oxford: OUP.
Lewis, D. (1979) 'Attitudes *De Dicto* and *De Se*', *Philosophical Review*, 88: 513–43.
Longworth, G. (2013) 'Sharing Thoughts about Oneself', *Proceedings of the Aristotelian Society*, 113: 57–81.
Lowe, E. J. (1986) 'What Do We See Directly?', *American Philosophical Quarterly*, 23: 277–85.
Lowe, E. J. (2008) 'Illusions and Hallucinations as Evidence for Sense', in E. Wright (ed.) *The Case for Qualia*, pp. 59–72. Cambridge, MA: MIT Press.
Magidor, O. (2015) 'The Myth of the *De Se*', *Philosophical Perspectives*, 29: 249–83.
McGinn, C. (2004) *Mindsight: Image, Dream, Meaning*. Cambridge, MA: Harvard University Press.
McGlynn, A. (2021) 'Immunity to Wh-misidentification', *Synthese*, 199: 2293–313.
Michaelson, E. (2016) 'The Lying Test', *Mind and Language*, 31: 470–99.
Millikan, R. (1990) 'The Myth of the Essential Indexical', *Noûs*, 24: 723–34.
Morgan, D. (2019) 'Thinking about the Body as Subject', *Canadian Journal of Philosophy*, 49: 435–57.
Mullins, S. and Spence, S. (2003) 'Re-Examining Thought Insertion', *British Journal of Psychiatry*, 182: 293–8.
Nagel, T. (1974) 'What Is It like to Be a Bat?', *Philosophical Review*, 83: 435–50.
O'Brien, L. (2007) *Self-Knowing Agents*. Oxford: OUP.
Palmira, M. (2020) 'Immunity, Thought Insertion, and the First-person Concept', *Philosophical Studies*, 177: 3833–60.
Palmira, M. (2022) 'Questions of Reference and the Reflexivity of First-Person Thought', *Journal of Philosophy*, 119: 628–40.
Peacocke, C. (1983) *Sense and Content: Experience, Thought, and Their Relations*. Oxford: Clarendon Press.
Peacocke, C. (2008) *Truly Understood*. Oxford: OUP.
Peacocke, C. (2012) 'Explaining *De Se* Phenomena', in S. Prosser and F. Recanati (eds) *Immunity to Error through Misidentification: New Essays*, pp. 144–57. Cambridge: CUP.
Peacocke, C. (2021) 'Self and Self- Representation in the Long Twentieth Century. A Critical Discussion', in P. Kitcher (ed.) *The Self. A History*, pp. 295–16. Oxford: OUP.
Pelczar, M. (2019) 'Defending Phenomenalism', *The Philosophical Quarterly*, 69: 574–97.
Perky, C. W. (1910) 'An Experimental Study of Imagination', *American Journal of Psychology*, 21: 422–52.
Perry, J. (1979) 'The Problem of the Essential Indexical', *Noûs*, 13: 3–21. Also in his *The Problem of the Essential Indexical and other Essays*. Oxford: Oxford U. P., 1993, 33–50, from which I quote.
Price, H. H. (1932) *Perception*. London: George Allen and Unwin.
Prosser, S. (2012) 'Sources of Immunity to Error through Misidentification', in S. Prosser and F. Recanati (eds) *Immunity to Error through Misidentification: New Essays*, pp. 158–79. Cambridge: CUP.
Quilty-Dunn, J. (2019) 'Unconscious Perception and Phenomenal Coherence', *Analysis*, 79: 461–9. https://doi.org/10.1093/analys/any022
Raleigh, T. (2019) 'Introduction: The Recent Renaissance of Acquaintance', in J. Knowled and T. Raleigh (eds.) *Acquaintance*, pp. 1–30. Oxford: OUP.
Recanati, F. (2016) *Mental Files in Flux*. Oxford: OUP.
Sainsbury, M. (2011) 'English Speakers Should Use 'I' to Refer to Themselves', in A. Hatzimoysis (ed.) *Self-Knowledge*, pp. 212–42. Oxford: OUP.
Salje, L. (2019) 'The Essential Non-Indexical', *Philosophers' Imprint*, 19: 1–13.

Salje, L. (2020) 'Lit from Within: First-Person Thought and Illusions of Transcendence', *Canadian Journal of Philosophy*, 50: 735–49.
Salje, L. (2024) 'Remember Me? First Person Thought, Memory and Explanations of IEM', *Philosophical Psychology*, Online First. https://doi.org/10.1080/09515089.2024.2386150
Saul, J. (2012) *Lying, Misleading, and What Is Said*. Oxford: OUP.
Schear, J. K. (2009) 'Experience and Self-Consciousness', *Philosophical Studies*, 144: 95–105.
Searle, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: CUP.
Seeger, M. (2015a) 'Immunity and Self-Awareness', *Philosophers' Imprint*, 15: 1–19.
Seeger, M. (2015b) 'Authorship of Thoughts in Thought Insertion: What Is It for a Thought to be One's Own?', *Philosophical Psychology*, 28: 837–55.
Sellars, W. (1963) 'Empiricism and the Philosophy of Mind', in *Science, Perception and Reality*, pp. 127–96. London: Routledge and Kegan Paul.
Stainton, R. (2019) 'Re-reading Anscombe on 'I'', *Canadian Journal of Philosophy*, 49: 70–93.
Stecker, R. (2013) 'Film Narration, Imaginative Seeing, and Seeing-In', *Projections: The Journal for Movies and Mind*, 7 147–54.
Stoljar, D. (2018) 'Introspection and Necessity', *Noûs*, 52: 389–410.
Stoljar, D. (2023) 'Is There a Persuasive Argument for an Inner Awareness Theory of Consciousness?', *Erkenntnis*, 88: 1555–75.
Terrone, E. (2020) 'Imagination and Perception in Film Experience', *Ergo*, 7: 161–89.
Torre, S. (2018) 'In Defense of *De Se* Contents', *Philosopy and Phenomenological Research*, 97: 172–89.
Verdejo, V. (2018) 'Thought Sharing, Communication and Perspectives about the Self', *Dialectica*, 72: 487–507.
Williamson, T. (1996) 'Cognitive Homelessness', *Journal of Philosophy*, 93: 554–73.
Williford, K. (2015) 'Representationalisms, Subjective Character, and Self-Acquaintance', in T. Metzinger & J. M. Windt (eds) *Open MIND*, 39(T). Frankfurt am Main: MIND Group.
Wiseman, R. (2017) 'What Am I and What Am I Doing?', *Journal of Philosophy*, 114: 536–50.
Wiseman, R. (2019) 'The Misidentification of Immunity to Error through Misidentification', *Journal of Philosophy*, 116: 663–77.
Wollheim, R. (1998) 'On Pictorial Representation', *The Journal of Aesthetics and Art Criticism*, 56 217–26.
Wright, C. (1998) 'Self-Knowledge: The Wittgensteinian Legacy' in C. Wright, B. C. Smith and C. McDonald (eds) *Knowing Our Own Minds*, pp. 13–45. Oxford: Clarendon Press.
Wu, W. (2012) 'Explaining Schizophrenia: Auditory Verbal Hallucination and Self-Monitoring', *Mind & Language*, 27: 86–107.
Zahavi, D. (2005) *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.
Zahavi, D. (2018) 'Consciousness, Self-Consciousness, Selfhood: A Reply to Some Critics', *Review of Philosophy and Psychology*, 9: 703–18.