

JOURNAL PRE-PROOF

Thoughts about oneself to share in context: meeting Bermúdez's challenge

Víctor M. Verdejo

DOI: 10.1387/theoria.26297

Received: 08/05/2024

Final version: 04/06/2025



This is a manuscript accepted for publication in *THEORIA. An International Journal for Theory, History and Foundations of Science*. Please note that this version will undergo additional copyediting and typesetting during the production process.

THOUGHTS ABOUT ONESELF TO SHARE IN CONTEXT: MEETING BERMÚDEZ'S CHALLENGE

Pensamientos sobre uno mismo para compartir en contexto: una respuesta al reto de Bermúdez

Víctor M. Verdejo
Pompeu Fabra University

ABSTRACT: Suppose you utter the sentence “I am a professional philosopher”. Can I—or anybody else—literally express the same thought you thereby expressed? An affirmative answer implies a potential split between the referent of the thought you expressed and its thinker, as well as the possibility of expressing that thought without using the first person pronoun. Here I attempt to clarify the basic features of a reference rule individuating such an intersubjectively shareable type of thought, i.e. the self type. Doing so requires meeting Bermúdez’s (2019) challenge, namely, the challenge of determining (a) in virtue of what is an individual the referent but not necessarily the thinker of the thought in such cases, and (b) what makes it the case that a sentence is used to express a self type of thought, as opposed to a thought of another type. I advance an elucidation that is entirely congenial to Bermúdez’s own account in terms of location abilities. Contrary to what he suggests, however, the analysis plausibly remains one at the level of the contextually constrained type (or instantiable type), and leads to a revision – if not a straightforward rejection – of Frege’s Criterion for the individuation of thought.

Keywords: Bermúdez, self, Frege’s Criterion, reference rule, thought sharing

RESUMEN: Supongamos que profieres la oración “Soy un filósofo profesional”. ¿Puedo yo —o cualquier otra persona— expresar literalmente el mismo pensamiento que tú expresaste al hacerlo? Una respuesta afirmativa implica una posible separación entre el referente del pensamiento expresado y su pensador, así como la posibilidad de expresar dicho pensamiento sin utilizar una forma de primera persona. En este trabajo, intento aclarar las características básicas de una regla de referencia que individúe este tipo de pensamiento intersubjetivamente compartible, es decir, el tipo “self”. Para ello, es preciso encarar el desafío planteado por Bermúdez (2019), a saber: (a) en virtud de qué es un individuo el referente, pero no necesariamente el pensador, del pensamiento en estos casos, y (b) qué determina que una oración se utilice para expresar un pensamiento del tipo “self” en lugar de un pensamiento de otro tipo. Propongo una elucidación que resulta plenamente compatible con la explicación de Bermúdez en términos de habilidades de localización. Sin embargo, y en contra de lo que él sugiere, este análisis sigue

situándose plausiblemente en el nivel del tipo contextualizado (o tipo instanciable), y conduce a una revisión —si no a un rechazo directo— del Criterio de Frege para la individuación del pensamiento.

Palabras clave: Bermúdez, yo, Criterio de Frege, regla de referencia, compartir pensamientos

SHORT SUMMARY: Can different people express the same thought in interactions with “I” and “you”? This paper articulates a positive answer, highlighting the basic features of a reference rule for such thoughts, addressing an important challenge posed by José Luis Bermúdez, and revisiting the role of context and Frege’s Criterion in the individuation of thought.

1. Introduction

The dispute over the alleged shareability of first person or ‘I’-thought has roots stretching back to Frege’s developments (Frege 1956). Taking sides in this dispute, and although specific articulations obviously vary, a number of authors would concur with the view that different people may express or grasp the same first person or ‘I’-thought in a given context.¹ José L. Bermúdez (2005, 2017, 2019) would further suggest that this constitutes, in fact, a fundamental constraint in the account of the concept or sense of ‘I’ which he terms the “Symmetry Constraint”.

I do not presently attempt to contribute to the assessment of this view, which is supported by a number of considerations, such as reflection on the nature of same-saying and disagreement (Bermúdez 2017, chap. 4; Verdejo 2018b), and is furthermore arguably consistent with a Fregean approach (Bermúdez 2005; Harcourt 1999; May 2006; Verdejo 2019) – see however Valente (2025) for discussion.² Here, I shall confine attention to the challenge that Bermúdez (2019) has posed to a particular articulation of this view based on the statement of reference rules for the first person (Verdejo 2018a, 2018b, 2021a). The challenge affects, first of all, those views ready to accept the intersubjective shareability of the type of thought expressible with utterances of sentences containing the first person pronoun, the ‘self type’ for short.³ Such an intersubjective shareability implies a potential split between the referent of an instance of

¹ Apart from the ones in the main text, they include Harcourt (1999), Longworth (2013, 2014), May (2006), Rödl (2007), Travis (2013), Skiba (2017) or Howell (2023).

² Valente (2025) critically revises the reasons in favor of the Symmetry Constraint and makes a forceful case that it clashes with the fundamental requirements of action explanation – what he calls the ‘Asymmetry Argument’. Valente’s insightful analysis deserves more space than I can provide here.

³ I use “type of thought” in a theoretically neutral way to refer to the kind or type of entity expressible and graspable via an utterance of a natural language sentence. I endorse no commitment to this entity being necessarily expressible or graspable, and may be considered independently of language. As we shall see (sect. 2), type analysis in the case of the self type is more complex than in other cases, but I don’t take type analysis *per se* to yield a commitment to any particular theoretical individuation of the type of entity in question. I thank an anonymous referee for helpful discussion.

this type of thought and its thinker, as well as the possibility of expressing an instance of that type with a sentence not featuring the first person pronoun, what we may call a ‘non-first-personal sentence’.

Suppose, for instance, that you wish to make conversation with someone on the bus home and utter (i):

- (i) I am a professional philosopher.

Suppose further that, having understood what you said, your (perhaps incredulous) interlocutor utters (ii) in reaction:

- (ii) You are a professional philosopher.

In this kind of setting, expression of the same thought as an utterance of (i) by someone different from you requires that the thinker of that thought be different from you, and also that it be expressed via an utterance of a non-first-personal sentence, that is to say, a sentence such as (ii) containing a singular term that refers to you, but is not the first person pronoun.

These niceties turn out to lead to remarkable complexity in the satisfactory account of the type of thought involved, viz. the self type. Bermúdez’s challenge concerns specifically those approaches which would wish to individuate this type – and indeed any type of thought – by means of an appropriate reference rule.⁴ According to Bermúdez (2019, 630-635), such an account might allow for shareability, as a reference rule for the self type can be devised to distinguish the referent (or as Bermúdez suggests, the subject) and the thinker of an instance of the thought. Nonetheless, the account would be at fault without further determining (a) in virtue of what is an individual the referent but not necessarily the thinker of the self type of thought expressed with a non-first-personal sentence? To put it another way, what can enable a self type of thought to refer to an individual different from its thinker? And (b) what makes it the case that such a non-first-personal sentence would express a self type of thought, as opposed to a thought of a different type? Or, in other words, what are the reasons to suppose that a thought of precisely the self type is being expressed in these cases, as opposed to, say, a thought of a demonstrative or descriptive type?

I here advance an elucidation of the basic features of a token-reflexive reference rule individuating the shareable self type to specifically respond to Bermúdez’s important challenge. In the course of doing so, I shall also examine the question of what is the correct level of analysis for the characterization of the self type and, relatedly, whether the resulting view is compatible with a baseline principle for individuating thought in terms of cognitive significance, namely, the so-called ‘Frege’s Criterion’ or ‘Frege’s Constraint’. I shall contend that the account offered here is entirely congenial to Bermúdez’s own account in terms of location abilities. However, contrary to what he suggests (Bermúdez 2019), the analysis plausibly

⁴ These are what Coliva & Palmira (forthcoming) and others aptly call “metasemantic” accounts.

remains one at the level of the contextually constrained type (or ‘instantiable type’), and leads to a revision – if not a straightforward rejection – of Frege’s Criterion for the individuation of thought.

2. *The instantiable self type of thought and Bermúdez’s challenge*

Let us, first off, spell out Bermúdez’s challenge in more detail. The discussion in which it is laid out is premised on SYMMETRY and COGNITIVE:

(SYMMETRY)

The thought expressible with the first person pronoun ‘I’ is interpersonally shareable.

(COGNITIVE)

Thoughts are individuated at the level of cognitive significance.

Coming up with a satisfactory characterization of the self type in accordance with SYMMETRY (labelled after Bermúdez’s Symmetry Constraint) and COGNITIVE constitutes no small task. One key question concerns the level or levels at which such a characterization is to be given. Following an earlier analysis (Verdejo 2018a, §1), here it is crucial to discriminate three ways in which an episode or event of thinking a thought expressible with the first person can be regarded: the level of the instantiated thought, the instantiable type of thought, and the general type.

Consider (i) again. When you *and* I express a thought via an utterance of (i), our thoughts do not even refer to the same person and can only be the same in the sense of being of the same general, overarching type: the type associated with uses of ‘I’ (type level). By contrast, when you utter (i) twice over, not only are the thoughts expressed uncontroversially the same from the point of view of the general type, they are also the same from the point of view of the type as contextually constrained or indexed with a particular subject as its referent, namely, yourself (instantiable type). Finally, the very instantiation of a thought through an utterance of (i) cannot be shared at all: these instantiations are unique unrepeatable spatiotemporal episodes of thinking.⁵

Arguably, only the instantiable type (i.e. the type as constrained by explanatorily salient instantiable features in the context) seems significant in discussions of the *interpersonal* shareability of thought expressible with the first person (Verdejo 2018a). To see this, take now (ii) above and let us ask: can a thought expressible with (i) by a subject A ever be the same as the thought expressible with (ii) by a distinct subject B? To answer affirmatively to this question by simply invoking the general (personal

⁵ On Davidson’s conception, they are unrepeatable, dated mental events (Davidson 1970) but there is no commitment here to a model on which individual thoughts are themselves just events, and not objects or contents of events of thinking. I thank an anonymous reviewer for prompting this clarification.

pronoun) type of thought seems quite correct but hardly illuminating. But so is the idea that the thoughts are not shareable in the sense that they involve distinct and unique instantiated episodes of thinking. The heart of the matter is therefore whether two different people may ever share an instantiable type of thought: in particular, whether the instantiable type of thought expressible with (i) – i.e. the general type as constrained by the utterer of (i) in the context as its referent – can ever be the same as the instantiable type of thought expressible with (ii) – i.e. the general type as constrained by the addressee of an utterance of (ii) in the context as its referent. I have suggested (Verdejo 2018a, §4) that shareability can be attained in these cases if the account of instantiable type forgoes the requirement that only the referent of the thought expressible with (i) can be its thinker.⁶ This evinces the need to enrich the token-reflexive reference rule individuating the self type to include at least two components or conditions which may or may not correspond to the same entity in a particular instantiation: the thinker of the thought and its referent.

Bermúdez (2019) agrees only partially with this diagnosis. While he believes that instantiable types can capture one sense of interpersonal shareability that is of interest in its own right, he does not think this level offers all that is needed in the account of the shareability of thoughts expressible with (i) and (ii). Suppose we accept a token-reflexive reference rule that fixes the referent of the thoughts in terms of “the subject of the thought”, as opposed to “the thinker”. This, Bermúdez contends, can hardly be the whole story. He writes:

So, on this formulation, a thought falls under the first person [self] type iff its referent is fixed by its subject. In many cases the thinker of the thought and the subject of the thought will coincide, but in some cases (including (ii), according to Verdejo), a thought can have a subject distinct from its thinker. (Bermúdez 2019, 634)

This understanding of the self type, Bermúdez observes, leads to a challenge:

But now the challenge for Verdejo’s account should be clear. It is all very well to define a version of the token-reflexive rule that applies to the subject of the thought, rather than to its thinker, but the key question now becomes: What makes an individual the subject of a first person thought in those cases where the subject of the thought is distinct from its thinker? [...] in virtue of what is A the subject of the first person thought that B expresses with (i)? It is not hard to explain why A would be the subject of a *second* person thought that B might express with [(ii)]. But Verdejo needs to explain how (ii) can express a *first* person thought. Moreover, Verdejo also needs to explain what makes it the case

⁶ This requirement is built into what I called ‘Thinker Distinctness’ (Verdejo 2018a, 580), namely, the claim that a difference in thinker of a thought expressed with ‘I’ entails a difference in the thoughts thereby expressed. A reviewer fairly asks: could further type distinctions be considered here? This discussion is premised on a negative answer, with perhaps the exception of Bermúdez’s more contextually constrained ‘token-sense’, which I will examine later (sects. 4 and 7). Note however that there is room for different (also shareable and non-shareable) articulations of instantiable type.

that sentences such as (ii) express second person thoughts in some contexts, and first person thoughts in others. (Bermúdez 2019, 634-635)

Bermúdez is clearly right that the challenge posed by these two pressing questions has not been properly addressed in the literature. To repeat: (a) in virtue of what is an individual the referent or subject, but not necessarily the thinker, of the self type of thought expressed with an utterance of a non-first-personal sentence such as “You are a professional philosopher”? In other words, what can enable a self type of thought to refer to something other than its thinker? And (b) what makes it the case that such a sentence expresses a thought of the self type – as opposed to a different, second person type – and does so in some contexts and not others? In what follows, I shall attempt to answer these questions. Accomplishing this task will contribute, I believe, to our understanding of the level of analysis relevant for the articulation of a shareable self type of thought in a context, and for the critical revision of the role of Frege’s Criterion in its individuation.

3. *Meeting Bermúdez’s challenge*

In this section, I shall develop the reference-rule account of first-person thought in order to meet Bermúdez’s challenge. What we need is a reference rule that individuates the self type of thought, namely, the type of thought paradigmatically but not exclusively expressible with the first person pronoun.⁷ We may build on Bermúdez’s own suggestion, and take that type to be individuated by the rule that for something to be the reference of one of its instantiations in thought is for it to be the subject of the thought so instantiated (cf. Bermúdez 2019, 634). In other words, the reference rule in question fixes the referent as the subject of a thought in a context. This statement of the rule echoes other influential statements in the literature (Campbell 1994, chap. 3; O’Brien 2007, chap. 4; Kripke 2011, 304). Let’s call this the *self rule*.

Much more could be said about the nature of subjects featuring in this reference rule, and the kinds of awareness of a subject sustaining instantiations of the self type. Various options would be open in the full elucidation of the condition “the subject” or “the person” of the thought in question. For instance, it is I think a common presupposition of Bermúdez’s and other proposals that a subject here refers to a self-conscious individual or an individual that is able to self-refer (e.g., Bermúdez 1998, §1.4; Sainsbury 2011; Echeverri 2020, §2). Another key feature here seems to be that this subject is the producer of the

⁷ Strictly speaking, the self type of thought is the type of thought that contains as a constituent the self type of concept expressible with e.g., the first person pronoun, and the self rule below will individuate that type of concept. I shall avoid this precisification in the main text for simplicity.

thought in question (e.g., Campbell 1994, 73; O'Brien 2007, 49).⁸ To move the discussion forward, I will here remain neutral regarding these refinements. On any of these construals, however, the self rule leaves open whether or not the subject acting as the referent of the thought in the context is, moreover, the thinker of the thought in question.

With the aid of the self rule, the answers to Bermúdez's key questions suggest themselves. In reply to (a), it seems plausible that the subject of a thought expressed by someone's uttering "You are a professional philosopher" is A in a context c iff A satisfies the referential condition specified by the self rule in c. A will be in this context the subject of the thought but not the thinker. Reluctance to accept this point might stem from suspicion that the only way in which the self rule – and presumably any rule involving psychological predicates – can be applied is through some sort of direct or introspective awareness of the thought only available to A. But we need not accept any such restriction, which would in fact be hard to reconcile with SYMMETRY. A subject B might be aware of a thought whose subject is A through indirect forms of awareness open to others (cf. Bermúdez 1998, chap. 9). In appropriate contexts, one reference rule – the self rule – will thus be the same for utterances of (i) and (ii) above. On this account, the thought expressible with the first person pronoun in (i) may indeed share a reference rule *and* a referent (i.e. a subject) with the thought expressible with the second person pronoun in (ii). Yet obviously the thoughts corresponding to (i) and (ii) will differ with respect to the individual playing the thinker role.

The relevant sense in which (i) and (ii) can be used to express the same thought concerns, to reiterate, the instantiable type of thought and not merely the overarching general type: i.e. it concerns the self type as indexed with A as its referent (see also the next sect. 4). What is crucial here is that reference to A is achieved via an application of the self rule. This is only one among several ways in which we may refer to A in thought. So understood, the self type must be sharply distinguished from linguistic type or Kaplanian character even if we consider this type at the highest level of generality. Arguably, (i) and (ii) belong to distinct linguistic types that can nonetheless be used to express thoughts of the same type, namely, the self type.

These considerations already put us in a position to answer the second of Bermúdez's questions (b). For they suggest that, in the relevant cases, there will be no such thing as a self type of thought – attached to (i) – distinct from a second person, demonstrative or descriptive type of thought – attached to (ii). By contrast, one and the same type of thought – a thought we employ to think about the subjects of episodes of thinking in a context – may be expressed by recourse to distinct first person or second person

⁸ This last suggestion, which emphasises the agentive aspect of the referent of the self type of thought, comports nicely with Bermúdez (2018) treatment of the sense of agency, but the argument in the text does not require it. See also Verdejo (forthcoming) for relevant discussion.

pronouns. From this angle, an utterance of (ii) in c expresses a self type of thought iff it involves the instantiation of a thought individuated by the self rule.

One may press on at this point and ask: what makes an utterance of (ii) referring to A be one that involves an application of the self rule as opposed to another rule individuating a co-referential thought in the context? For instance, why is that utterance not an application of, as it might be, the rule to refer to the salient professional philosopher in the context, or the rule to refer to one's addressee in the context? To appropriately answer this question we will need to attend to the particular explanatory contribution made by different proposals. First, the ascription of a shareable self type would need to be in harmony with the different phenomena traditionally associated with thought expressible with the first person pronoun, such as immunity to error through misidentification (Shoemaker 1968), and related properties such as guaranteed self-reference (Coliva 2003; Echeverri 2020; García-Carpintero forthcoming) or so-called 'illusions of transcendence' (Coliva 2012; Salje 2020). But there is no obvious reason to suppose that the self rule cannot accommodate these phenomena. In fact, for example, I have argued that the reference-rule articulation of the shareable self type may feature in explanations of immunity to error through misidentification that better capture its ground-relativity (Verdejo 2021b).

Moreover, one may consider the difference between the ascription of a self type of thought and the ascription of alternative types of concepts co-referential in the context. For instance, it might be argued that, to competently use the second person pronoun in (ii), B need not know what a professional philosopher actually is; nor would B seem to require to succeed in addressing A to refer to A in thought. One may likewise take into account the inferential pattern of thoughts expressible with "I" and "you". Assuming epistemically adequate circumstances, A may arguably be in a position to infer that something is F and G, from the thought expressed by A with "I am F" and the thought expressed by B with "You are G", without A needing to draw on an intermediate premise to the effect that the reference of "I" and "you" in the context is the same (cf. Campbell 1994, chap. 3). This is a distinctive inferential pattern that "I" as used by A would not plausibly exhibit on other second-personal, demonstrative or descriptive accounts of the thought expressible with "you".⁹ More generally, and granted that we accept the explanatory postulation of reference rules at all, the situation here does not seem to be different from other cases in which more than one possible candidate is contemplated. There must of course be a fact of the matter about which reference rule a particular thinker is applying in a context. And sometimes this might be hard to determine, given the nature of the case (idiosyncratic use, infelicitousness, vagueness...). But this doesn't seem to pose a problem specific to personal pronouns or the self rule.

⁹ I am grateful to an anonymous referee for highlighting the need to clarify these points.

4. *The instantiable self type and location abilities*

The answers just provided are not, as far as I can see, very far from the elucidation of thought sameness for (i) and (ii) that Bermúdez himself affords in terms of location abilities (Bermúdez 2011, 2017, 2019). According to Bermúdez, “the ability to locate the referent of the sentence that I exercise in using ‘I’ with understanding is the very same ability that is exercised when someone else grasps sentences that I utter using ‘I’”. It is also the same ability that you exercise when, in contexts like the one which we are considering with (i) and (ii), you think thoughts about me that you express with the second person pronoun” (Bermúdez 2019, 635-636).¹⁰

Importantly, Bermúdez also maintains that sameness of thought in these cases involves two cognitively different ways of exercising an ability associated with (i) and (ii) respectively. In Bermúdez’s words: “Of course, the way in which you go about determining whether I am F may be very different from how I go about determining whether I am F. The first-person and second-person ways of establishing F-ness may differ, but that has no bearing on the relation between your understanding of ‘I’ and my understanding of ‘I’, both of which exploit a single ability” (Bermúdez 2017, 116-117).

However, there are two crucial issues that need to be taken into account for a full assessment of Bermúdez’s position at this point. First, and contrary to what Bermúdez suggests, the articulation of sameness he provides is plausibly one that is found at the level of instantiable type in my terminology. This point dovetails, secondly, with a revision of the role of Frege’s Criterion in the analysis of these cases, a chief constraint in Bermúdez’s own approach. The first issue will occupy the remainder of this section. I will take up the second, deeper one in the next two (sects. 5 and 6).

Note, to begin with, that the considerations offered as solutions to Bermúdez’s challenge so far fall clearly beyond the general type of thought – as individuated by a given reference or token-reflexive rule – or what Bermúdez calls the type-sense (Bermúdez 2019, 635). Obviously too, the analysis does not belong either to the level of non-repeatable instantiated type which Bermúdez also agrees cannot itself be shared (ibid, 631). It all seems to come together if we take Bermúdez’s account therefore to also proceed at the level of instantiable, contextually constrained, type of thought.

Bermúdez disagrees. He takes his analysis to be “firmly pitched at the level of token-sense” instead (Bermúdez 2019, 637). According to Bermúdez, the token-sense of ‘I’ is to be understood “in terms of abilities to determine the truth-condition of token sentences involving ‘I’ – and in particular the ability to locate the referent of the sentence” (ibid, 635). So, on Bermúdez’s view, the question of whether – in

¹⁰ It might reasonably be asked how the account in terms of location abilities should be understood in special cases: e.g., those in which distinct subjects arguably share thoughts expressible with “I” but no location ability is in view (e.g., utterances of (i) and (ii) on the phone or a Zoom session, or when they are just written). I will not however press this line here.

which contexts – do utterances of (i) and (ii) by A and B, respectively, involve the same thought or not is, in fact, the question of whether such utterances involve the same ability to determine the truth-condition of the corresponding sentences – and the location in space of A. But this way of putting things does not contravene the idea that token-sense is a notion at the level of instantiable type.

The defining trait of an analysis at the level of instantiable type is that, in order to determine whether utterances of (i) and (ii) would significantly involve *the same thought*, we need to heed (both over the general type and instantiated type) the features that are salient or explanatorily relevant in the context (see sect. 2). These are given, on the reference-rule account, by the reference of the thoughts as fixed by the self rule in the context. According to Bermúdez, these are given by the abilities to locate a thinker exercised in the context. It is therefore quite plausible that, as anticipated above, the debate here is properly located at the level of the contextually constrained instantiable type.

The correct identification of the level at which the analysis of sameness operates obviously matters in its own right. Yet the present considerations have a crucial ramification regarding the role of Frege's Criterion in the picture. This will be the topic of the next sections.

5. *Symmetry and Frege's Criterion*

Frege's alleged criteria for either sameness or distinctness of thought may be exposed to challenging problems regarding their proper focus (Schellenberg 2012), or may be thought to involve no substantive or positive content (Speaks 2013), or may be based on an unpalatable form of circularity given the interference of natural language in the expression of thought (Textor 2018). Yet here I shall focus attention solely on the question of whether Frege's Criterion – interpreted as a sufficient condition for the distinctness of thought – is compatible with sameness of thought in 'I'/'you'-interactions or SYMMETRY above (sect. 2). So devised, the discussion will therefore not adjudicate the overall soundness of Frege's Criterion, or SYMMETRY for that matter, and will thus rest on their initial plausibility.

As noted, however, the present discussion also assumes COGNITIVE (sect. 2). The considered accounts of the (shareable) sense of 'I' thus factor in the distinctive cognitive significance that is associated with utterances of 'I' in the relevant cases. This means, for instance, that the sense expressible with 'I' must generally be distinguished from the sense expressible, say, with other potentially co-referring terms such as 'that professional philosopher' or 'José'. Importantly, however, COGNITIVE involves no straightforward commitment to Frege's Criterion. Put bluntly, Frege's Criterion is a (well-known) principle for spotting different thoughts in accordance with COGNITIVE, yet COGNITIVE could perfectly stand even if Frege's Criterion falls insofar as we have a positive story to tell about cognitive-significance based individuation. In fact, there is an initial motivation to suspect that, granted SYMMETRY, Frege's

Criterion must indeed fall in the light of the positive stories already sketched above. For to accede to sameness of thoughts expressible with 'I' and 'you' is, as we have seen, also to accede to different (first- and second-person) ways of exercising an ability (on Bermúdez's account), or different thinker roles operative in thinking one and the same self type of thought (on the reference-rule account). The tension is plain if we take Frege's Criterion to discriminate types of thought in terms of such disparate ways or roles.

However, contrary to what I have suggested (Verdejo 2018a, 582-583; see also Verdejo 2018b, 2021a), Bermúdez is surely right that there is no straightforward argument from an account of sameness at the level of instantiable type – or token-sense in Bermúdez's terminology – to the conclusion that Frege's Criterion is indeed violated. Frege's Criterion is a criterion that detects differences between thoughts. To a first approximation, we may take the criterion to state that subjects cannot coherently or rationally have conflicting attitudes towards the same thought:

Frege's Criterion

For any thoughts, 'I' and 'I', and subject S, if S can rationally believe (be agnostic about/reject) 'I' and be agnostic about/reject (believe) 'I', then $I \neq I$.

It just follows from this characterization that, in the contexts in which 'I' and 'I' are determined to be the same (e.g., because they are determined to involve the same ability to locate a self-referring thinker in space), no difference can be detected by the criterion. There is hardly anything objectionable with this basic point as far as it goes. But note that the point does not put us in a better position to explain either sameness of thought – and a fortiori sameness of thought expressible with personal pronouns – nor whether Frege's Criterion can be preserved at the end of the day. Bermúdez writes:

As I formulate it, Frege's Criterion holds that no two utterances can express the same token-sense if it is possible for a thinker rationally to take different attitudes to them, and by 'utterances' I explicitly mean utterance-tokens. So, I am not making a claim at the level of instantiable type. And I am certainly not committed to any claim stronger than that, *in some contexts*, an utterance containing "I" *can* have the same sense (express the same thought) as an utterance containing "you". (Bermúdez 2019, 637, emphasis in original; see also Bermúdez 2017, 118)

Let us bracket now the retraction of the level of instantiable type (see sect. 4 above and also the concluding sect. 7).¹¹ Bermúdez's formulation of Frege's Criterion departs from our initial one and can be stated as follows:

¹¹ Importantly, the line of argument laid out here and the next sect. 6 is independent of and does not presuppose the notion of instantiable type.

Frege's Criterion in Context

For any sentences, s and s' , thoughts T and T' , subject S , and context c , if S can coherently or rationally believe (be agnostic about/reject) T as expressed by an utterance of s and be agnostic about/reject (believe) T' as expressed by an utterance of s' in c , then $T \neq T'$.¹²

This formulation involves a relativization to sentences and contexts of utterance – or “utterance-tokens” as Bermúdez puts it. Still, the relevant question is a possibility question, namely: is it really *possible* for a thinker to rationally take different attitudes to co-referring¹³ thoughts (or token-senses) expressible with utterances of ‘ T ’ and ‘ you ’ (as in (i) and (ii) above) in a context?

If the answer is negative, the compatibility of SYMMETRY with Frege's Criterion is admittedly untouched. If the answer is positive, however, Frege's Criterion will not provide the correct individuation of thought at the level of shareable token-sense: Frege's Criterion will force us to conclude that thoughts expressible with ‘ T ’ and ‘ you ’ involve distinct token-senses or location abilities after all. So, as I see it, the only way to show that the picture is compatible with Frege's Criterion is by providing an argument to the effect that the answer to this question is indeed negative. To insist that we restrict attention to the contexts in which ‘ T ’ and ‘ you ’ do contribute to the expression of one and the same thought – or involve one and the same location ability – can hardly count as supporting just this verdict. Similarly, if the answer is “it depends on whether the thoughts are the same or not”, then the criterion would not seem to be of much use or involve a genuine constraint in determining whether two utterances do indeed express the same thought in the context or not.

Importantly, Frege's Criterion in Context leaves open whether there might be a context – or a collection thereof – in which no conflicting rational attitudes are actually allowed for the thoughts that are indeed regarded to be *distinct*. In a footnote indebted to Paul Benacerraf, Evans makes this explicit when he observes that the criterion only requires that someone can coherently endorse conflicting attitudes towards the thoughts but not anyone:

The thought expressed by ‘Hesperus is F ’ is different from the thought expressed by ‘Phosphorus is F ’ but it is not true that *anyone* who understands the two sentences can take different attitudes to them. For example, it is not true for someone who knows that Hesperus is Phosphorus. (Evans 1982, 19, n. 19, emphasis in original)

Thus, the fact that a subject in a context may not be allowed to rationally have conflicting attitudes towards the thought expressed by a pair of sentences – such as sentences of the form ‘Hesperus is F ’ and

¹² See Evans 1981, 18-19; Harcourt 1999, 331. Evans quotes Frege's letters to Russell and Jourdain (1980, 80 and 153) in support of a similar reading of the criterion. It is logically entailed by the stronger equipollence criterion for the sameness of thought (Frege 1979, 197-198), also alluded to by Evans.

¹³ The assumption is that failure of co-reference would eo ipso involve the possibility in question.

‘Phosphorus is F’ as understood by someone who knows that Hesperus is Phosphorus – is orthogonal to the question of whether the thoughts so expressed are distinct or not according to the criterion. For the same reason, the fact that there is a context in which one is not allowed to have conflicting attitudes towards ‘I’/‘you’-thoughts of the sort expressed by (i) and (ii) – the contexts targeted by Bermúdez’s Symmetry Constraint – is orthogonal to the question of whether Frege’s Criterion is ultimately respected. More may be ventured, for to grant – as Bermúdez does in the quoted passage above (Bermúdez 2019, 637) – that ‘I’/‘you’-thoughts as expressed with (i) and (ii) will cognitively differ in some contexts and not others, already indicates that Frege’s Criterion must prescribe a difference in the thoughts so expressed. We can confirm this possibility with many examples. For instance, we may conceive of David Kaplan, John Perry or certainly ourselves accepting the thought expressed with (i) while at the same time squarely rejecting or being agnostic about the thought expressed with (ii) when inadvertently referring with ‘you’ to themselves as appearing on a blurry TV screen.

6. *Symmetry and double contextualization*

I take the foregoing to establish that the compatibility of sameness of thought expressible with ‘I’ and ‘you’ – and hence SYMMETRY – and Frege’s Criterion is at the very least contentious, and arguably in tension with a standard (Evansian) interpretation of the latter. In this section, however, I would like to explore in more detail the prospects for a reading of the criterion that supports the compatibility claim.

It might be objected, for instance, that the clash just brought out between SYMMETRY and Frege’s Criterion stems from questionably assuming that there is a straightforward correspondence between co-referring utterances (or tokens) of a sentence type and thoughts expressed therewith in every context. The unwarranted assumption is, more specifically, that co-referring utterances of ‘I’-sentences, on the one hand, and ‘you’-sentences, on the other, guarantee the expression of the same thought type across contexts.

The compatibilist objection would not here exactly be levelled against the idea that Frege’s Criterion ought to be articulated in such a way that sameness of thought matches sameness of type of expression. For, in order for ‘I’/‘you’-thought sameness to make sense at all – and assuming that ‘I’ and ‘you’ are indeed different types of expression – it already has to be true that thought-type identity doesn’t require expression-type identity (see also sect. 3). The compatibilist invokes rather an interpretation of Frege’s Criterion such that sameness of reference for one and the same expression type (viz. the linguistic types corresponding to ‘I’ or ‘you’) does not necessarily yield sameness of thought type across contexts.

To see this, suppose that A utters (i) and (ii) in a context to refer to what is, unbeknownst to A, the same person (namely, A). Call this the ‘contrasting case’. An application of Frege’s Criterion to the contrasting

case, therefore, would yield the result that the thoughts expressed through utterances of (i) and (ii) are distinct. Compare this to a context in which (i), as used by A, and (ii), as used by B, are uttered to express thoughts which A knows to be about A. Call this the ‘target case’. Now, to block the inference from the application of Frege’s Criterion to the contrasting case to the conclusion that the thoughts expressed in the target case are also distinct, the compatibilist has to question that utterances of (i) and (ii) in the contrasting case and the target case involve the same thoughts. Only thus can it still be true that the ‘I’- and ‘you’-thoughts in the contrasting case are distinct, whereas the ‘I’- and ‘you’-thoughts in the target case are the same, compatibly with Frege’s Criterion.

It will not suffice to respond to this move by noting that the contrasting and target cases plausibly involve the same *standards of use* of the expressions (viz. ‘I’/‘you’) and, hence, that the individuation issued in the contrasting case would reasonably be taken to cover the target case, and in fact, any context in which those uses are considered to be also standard. For our objector may propose, more pressingly, that the contexts in which co-referring and standard uses of ‘I’/‘you’ express the same thoughts concern a type of thought (viz. the self type) that is to be distinguished from the type of thought expressed by co-referring and standard uses of other contexts – marked out by the criterion – in which ‘I’/‘you’ do make room for conflicting but rational attitudes.

This is, I suggest, a special case of a more general kind of strategy that the advocate of compatibility may pursue. The more general strategy consists of restricting the validity of the individuation delivered by Frege’s Criterion to a context or contexts – namely, those for which cognitive differences registered by the criterion are deemed to be relevant – and exclude it from others – namely, those for which it is independently plausible that the thoughts under scrutiny are in fact the same. More precisely, this amounts to a relativisation of the criterion not only to a context of application (c_A) – as we have done thus far – but also to what we may call a ‘context of individuation’ (c_I):

Frege’s Criterion with Double Contextualisation

For any sentences, s and s' , thoughts T and T' , subject S , and contexts c_A and c_I , if S can coherently or rationally believe (be agnostic about/reject) T as expressed by an utterance of s and be agnostic about/reject (believe) T' as expressed by an utterance of s' in c_A , then $T \neq T'$ in c_I .¹⁴

Now, while there is nothing incoherent in this strategy per se, it does little by way of preserving Frege’s Criterion as a substantive principle for the individuation of thought. Once we concede that the cognitive differences registered by means of the criterion may have an effect in some contexts and not others –

¹⁴ This is also a reading of Frege’s Criterion that can be endorsed to respond to what Valente calls “non-transitivity cases” or cases that purportedly show that SYMMETRY is incompatible with the intransitivity of communication or testimonial knowledge transmission (Valente 2024, §2; see also Valente & Onofri 2023; Goodman forthcoming). For reasons made clear in the main text, I agree with Valente that this kind of contextual relativisation is counterproductive.

and do so irrespective of co-reference, sameness of expression type or standard of language use (for ‘I’ and ‘you’) – we open the door to any thought-difference detected by means of the criterion being arguably irrelevant for the context at stake, and indeed for any context other than the context in which the difference is detected (that is, irrelevant whenever $c_A \neq c_I$).¹⁵

This outcome seems to me to leave the criterion fatally vulnerable to charges of *unprincipledness* and *idleness*. The criterion is now unprincipled because it can serve to cherry-pick contexts of individuation and may fail to support any principled conclusions over contexts that would involve linguistically kosher co-referential uses of the same terms as the context of application. The criterion is also idle because any difference detected can now be imputed to the peculiarity of a context or time of application – as opposed to the type of thought – with no effect besides that particular context or time and, a fortiori, no effect concerning the individuation of thought.

The force of these charges is perhaps striking once we see that they jeopardise very basic differences drawn by the criterion in non-indexical cases. If double contextualisation is called upon, such differences may be said to hold in some contexts but not others, or even only hold for one-off contexts or times.

To illustrate, consider the classic ‘Hesperus’/‘Phosphorus’ example and let us assume that the relevant contexts (of application and individuation) are just times. If double contextualisation holds, the fact that Hammurabi rationally endorses conflicting attitudes at t_1 with respect to thoughts expressed with ‘Hesperus is F’ and ‘Phosphorus is F’ is not enough to establish that the thoughts differ at a distinct time t_2 . For now it is open to a theorist to select – from among the times different from t_1 – those times where the registered difference is in force and those where it is not. Note, in particular, that – without further ado – it is not guaranteed that the thoughts will be distinct at t_2 even if the reference and the standards governing the use of ‘Hesperus’ and ‘Phosphorus’ at that time are the same as in t_1 . No doubt, one may advert at this point to other considerations that would make it independently plausible that, at t_2 , the ‘Hesperus’/‘Phosphorus’-thoughts are distinct, or indeed the same. But since these considerations would admittedly go beyond what Frege’s Criterion says, the criterion does not itself really provide a positive, principled basis for the discrimination of thought expressible with ‘Hesperus’ and ‘Phosphorus’.

In consequence, it is only at t_1 that we may have the right to claim that thoughts expressible with ‘Hesperus’ and ‘Phosphorus’ are guaranteed to be distinct according to the criterion. But to concede such stark contextualisation backfires into idleness. For if Hammurabi becomes aware of the fact that ‘Hesperus’ and ‘Phosphorus’ co-refer at a later time t_2 it would cease to be true, according to the criterion

¹⁵ I take this point to apply, mutatis mutandis, to the articulation of Frege’s Criterion in terms of actual, as opposed to absolute, possibility advocated by Recanati (forthcoming). Detailed discussion of Recanati’s views in this connection must be left for another occasion.

so understood, that ‘Hesperus’/‘Phosphorus’-thoughts are distinct. For at t_2 , and contrary to Evans’s insight highlighted above, thought distinctness will not persist if – as a consequence of newly acquired knowledge – it ceases to be rational for Hammurabi to stand in conflicting attitudes towards thoughts expressed with ‘Hesperus’ and ‘Phosphorus’. In sum, if we restrict difference of thought to t_1 , the criterion turns out to be idle as a principle of individuation: it would at most reflect the specificities of t_1 but not the defining features of ‘Hesperus’/‘Phosphorus’-thoughts as such. In general, if we restrict the context of individuation to the context of application, even paradigmatic distinctions delivered by the criterion in non-indexical cases would run the risk of being idle for the individuation of thought.

One final move may be invoked in order to resist this result. Some authors have proposed to split the notion of a thought expressed by a sentence in cases of ‘I’/‘you’-interaction and perhaps elsewhere. Uses of (i) and (ii), the basic idea goes, may involve the expression of more than one thought or thought-content.¹⁶ If so, when Frege’s Criterion issues a difference regarding (i) and (ii), it might be taken to concern only one (class) of such thoughts, but precisely not those for which sameness is proposed. However, this objection misses its target. For to fracture the domain of thoughts associated with ‘I’/‘you’ between those in line with Frege’s Criterion and those that are the same is just to admit that one cannot, as a matter of fact, hold on to the criterion *and* SYMMETRY for the same (class of) thoughts.

I conclude that, while Frege’s Criterion is, in the ways we have earlier seen, uncontroversially contextually constrained with respect to its application, it can only be contextually constrained with respect to the individuation it yields at the risk of undermining the object of the criterion as a substantive individuating principle for thought.

7. *Going full circle: instantiable types again*

It must be noted that, in subscribing to both SYMMETRY and COGNITIVE, the views discussed here converge much more than they diverge from each other. That said, the root problem grappled with in these pages has ultimately to do, it seems to me, with the role of context in a satisfactory account that wishes to preserve both assumptions. Bermúdez takes himself to be operating at a highly contextually limited level (the ‘token-sense’ or the ‘token-utterance’) which is different from the repeatable instantiable type. But such a contextually limited approach, now we can perhaps see, does not involve much of a commitment regarding general principles for the individuation of thought. And this is why, Bermúdez correctly insists, the Symmetry Constraint – SYMMETRY – should only be regarded as operative “*in some contexts*”. But this, I hope to have shown, is problematic concerning the satisfactory characterization of

¹⁶ This sort of suggestion has been recently developed in Howell 2023, chap. 6. See also Chalmers 2011, 602 or Soames 2015, 112-116.

the shareable self type of thought and the substance of Frege's Criterion as tenets that are cross-contextually explanatory.

A consequence of the foregoing discussion is thus that the level of the instantiable type is plausibly the centre stage of discussions of shareability of the self type – a type of thought expressible with the first and second person pronouns. Accounts at that level can, when properly fleshed out in terms of reference rules, meet the important challenge posed by Bermúdez and do so in a way that is entirely congenial to his own account in terms of location abilities (see sect. 4). But even if we agree with this conclusion, there are also deeper issues to which these considerations direct us. They include whether these competing accounts adequately capture the role of context in the articulation of shareable thought and, rather crucially, whether Frege's Criterion may indeed be meaningfully preserved once distinct cognitive ways of exercising an ability or thinker roles (associated with 'I' and 'you' respectively) are posited in the account of the shareability of instantiable type, or token-sense.

Acknowledgements

This paper is fundamentally indebted to José Luis Bermúdez for his inspiring work and always insightful and constructive feedback. Earlier versions of this material were presented at the 30th SIUCC Conference (Valencia, 2023), and the 3rd Context, Cognition and Communication Conference (Warsaw, 2022). I am very grateful to the Logos Research Group, and the audiences and to two anonymous referees for their helpful comments. My special thanks to Valen Simpson, Matheus Valente, and Anna Verdet for stimulating discussion on these issues. This work has been supported by grants RYC2021-033972-I funded by MCIU/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR, and PID2023-151071NB-I00 funded by MCIU/AEI/10.13039/501100011033/FEDER, UE; as well as the Fundación Ramón Areces, project "Communication, Inner Speech and Mental Health" (CIM).

References

- Bermúdez, J.L. (1998). *The paradox of self-consciousness*. Cambridge, MA: MIT Press.
- Bermúdez, J.L. (2005). Evans and the sense of "I". In J.L. Bermúdez (Ed.). *Thought, reference, and experience: Themes from the philosophy of Gareth Evans* (pp. 164-194). Oxford: Clarendon Press.
- Bermúdez, J.L. (2011). Self-knowledge and the Sense of "I". In A. Hatzimoysis (Ed.). *Self-knowledge* (pp. 236-245). Oxford: Oxford University Press.
- Bermúdez, J.L. (2017). *Understanding "I": Language and thought*. Oxford: Oxford University Press.
- Bermúdez, J.L. (2018). First person awareness of agency. *Teorema*, 37(3), 21-38.

- Bermúdez, J. L. (2019). First person thoughts: Shareability and symmetry. *Grazer Philosophische Studien*, 96(4), 629-638.
- Campbell, J. (1994). *Past, space, and self*. Oxford: Clarendon Press.
- Chalmers, D. (2011). Propositions and attitude ascriptions: A Fregean account. *Noûs*, 45(4), 595-639.
- 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(8), 416-431.
- 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 & F. Recanati (Eds.). *Immunity to error through misidentification: New essays* (pp. 22-45). Cambridge: Cambridge University Press.
- Coliva, A. & Palmira, M. (forthcoming). Immunity to error through misidentification: some trends. *Philosophical Psychology*. doi: 10.1080/09515089.2024.2371423
- Davidson, D. (1970). Mental events. Reprinted in D. Davidson (Ed.). *Essays on action and events* (pp. 207-224). Oxford: Clarendon Press, 1980.
- Echeverri, S. (2020). Guarantee and reflexivity. *Journal of Philosophy*, 117(9), 473-500.
- Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.
- Frege, G. (1956). The thought: A logical inquiry. Translated by A.M. and M. Quinto. *Mind*, 65(259), 289-311.
- Frege, G. (1979). *Posthumous writings*. Translated by P. Long and R. White. Chicago: University of Chicago Press.
- Frege, G. (1980). *Philosophical and mathematical correspondence*. Abridged from the German edition by Brian McGuinness. Translated by Hans Kaal. Oxford: Basil Blackwell.
- García-Carpintero, M. (forthcoming). The *real guarantee* in *de se* thought: How to characterize it? *Philosophical Quarterly*. doi: 10.1093/pq/pqae133
- Goodman, R. (forthcoming). Shared thought and communication. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity*. Oxford: Oxford University Press.
- Harcourt, E. (1999). Frege on "I", "now", "today" and some other linguistic devices. *Synthese*, 121(3), 329-356.
- Howell, R.J. (2023). *Self-awareness and the elusive subject*. Oxford: Oxford University Press.
- Kripke, S. (2011). The first person. In S. Kripke (Ed.). *Philosophical troubles, volume 1* (pp. 292-321). New York: Oxford University Press.
- Longworth, G. (2013). Sharing thoughts about oneself. *Proceedings of the Aristotelian Society*, 113(1), 57-81.
- Longworth, G. (2014). You and me. *Philosophical Explorations*, 17(3), 289-303.
- O'Brien, L. (2007). *Self-knowing agents*. Oxford: Oxford University Press.
- May, R. (2006). Frege on indexicals. *Philosophical Review*, 115(4), 487-516.

- Recanati, F. (forthcoming). Sameness of mode of presentation. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity*. Oxford: Oxford University Press.
- Sainsbury, M. (2011). English speakers should use 'I' to refer to themselves. In A. Hatzimoysis (Ed.). *Self-knowledge* (pp. 246-260). New York: Oxford University Press.
- Salje, L. (2020). Lit from within: First-person thought and illusions of transcendence. *Canadian Journal of Philosophy*, 50(6), 735-749.
- Schellenberg, S. (2012). Sameness of Fregean sense. *Synthese*, 189(1), 163-175.
- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65(19), 555-567.
- Skiba, L. (2017). Frege's unthinkable thoughts. *Proceedings of the Aristotelian Society*, 117(3), 333-343.
- Soames, S. (2015) *Rethinking language, mind, and meaning*. Princeton (NJ): Princeton University Press.
- Speaks, J. (2013). Individuating Fregean sense. *Canadian Journal of Philosophy*, 43(5), 634-654.
- Textor, M. (2018). Frege's recognition criterion for thoughts and its problems. *Synthese*, 195(6), 2677-2696.
- Travis, C. (2013). *Perception: Essays after Frege*. Oxford: Oxford University Press.
- Valente, M. (2025). Asymmetries between 'you' and 'I'. *THEORIA. An International Journal for Theory, History and Foundations of Science*. <https://doi.org/10.1387/theoria.26508>
- Valente, M. & Onofri, A. (2023). A puzzle about communication. *Review of Philosophy and Psychology*, 14(3), 1035-1054.
- Verdejo, V.M. (2018a). On having the same first person thought. *Grazer Philosophische Studien*, 95(4), 566-587.
- Verdejo, V.M. (2018b). Thought, communication and perspectives about the self. *Dialectica*, 72(4), 487-507.
- Verdejo, V.M. (2019). In defence of the shareability of Fregean self-thought. *Acta Analytica*, 34(3), 281-299.
- Verdejo, V.M. (2021a). The second person perspective. *Erkenntnis*, 86(6), 1693-1711.
- Verdejo, V.M. (2021b). Perspectives on *de se* immunity. *Synthese*, 198(10), 10089-10107.
- Verdejo, V.M. (forthcoming). On the rationality of thought insertion judgments. *Philosophical Psychology*. doi: 10.1080/09515089.2023.2296580

Víctor M. Verdejo is a Ramón y Cajal fellow based in the Department of Humanities at Pompeu Fabra University (Barcelona). He is senior member of Logos Research Group and collaborator of the Valencia Philosophy Lab. Mainly interested in the nature of thought and concepts, and related areas of philosophy of language and cognitive science, his latest research focuses on communication, thought sharing, and the self.

Address: Departament d'Humanitats, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27,
08005 Barcelona, Spain. E-mail: victor.verdejo@upf.edu – ORCID: 0000-0002-7872-0206

