Theoria, 2026, 41(2), 283-338
https://doi.org/10.1387/theoria.28723

THEORIA
ISSN 0495-4548 – eISSN 2171-679X

Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako

(Yo, habla interna, racionalidad y explicación: respuestas a Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués y Jorba, Dohrn, Fisher y Jurjako)

José Luis Bermúdez*
Texas A&M University
https://orcid.org/0000-0002-8312-1013

Keywords

J.L. Bermúdez
Self
Inner speech
Rationality
Frames
Sense data
Free energy principle

ABSTRACT: This paper collects the responses by J.L. Bermúdez to articles discussing his work by Matheus Valente, Víctor Verdejo, Derek H. Brown, Shaun Gallagher, Carlota Serrahima, Daphne Bernués and Marta Jorba, Daniel Dohrn, Sarah A. Fisher and Marko Jurjako.

Palabras clave

J.L. Bermúdez
Yo
Habla interna
Racionalidad
Marcos
Datos sensoriales
Principio de energía libre

RESUMEN: Este artículo reúne las respuestas de J.L. Bermúdez a trabajos sobre su pensamiento escritos por Matheus Valente, Víctor Verdejo, Derek H. Brown, Shaun Gallagher, Carlota Serrahima, Daphne Bernués y Marta Jorba, Daniel Dohrn, Sarah A. Fisher y Marko Jurjako.

Gako-hitzak

J.L. Bermúdez
Nia
Barne-hizketa
Arrazionaltasuna
Markoak
Zentzumen-datuak
Energia askearen printzipioa

LABURPENA: Artikulu honetan, J L Bermúdezek bere pentsamenduari buruz hainbat egilek idatzitako lanei emandako erantzunak biltzen dira, egileok hauek izanik: Matheus Valente, Víctor Verdejo, Derek Brown, Shaun Gallagher, Carlota Serrahima, Daphne Bernués eta Marta Jorba, Daniel Dohrn, Sarah Fisher eta Marko Jurjako.

* ***Correspondence to:*** José Luis Bermúdez. Suite 409, Academic Building, Texas A&M University 77843 and Department of Philosophy, Texas A&M University, College Station TX 77843-4223. – jbermudez@tamu.edu – https://orcid.org/0000-0002-8312-1013

How to cite: Bermúdez, José Luis (2026). «Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 41(2), 283-338. (https://doi.org/10.1387/theoria.28723).

Received: 22 April, 2026; Final version: 19 June, 2026.
ISSN 0495-4548 - eISSN 2171-679X /

It has been an honor and a pleasure to write the responses collected together in this long paper – as indeed it was to attend the XXX Inter-University Workshop on Philosophy and Cognitive Science in 2023, from which many of the target papers are derived. The contributors to this special issue of *Theoria* have without exception produced thought-provoking and incisive discussions of my work in a number of different areas. Rather than respond to each of the papers individually, I decided to organize my responses thematically, as reflected in the section titles below:

Each of the sections is self-contained, and so nothing further is required by way of introduction, except to thank the editors of the special issue, Victor Verdejo and Matheus Valente for gathering together and editing such a wonderful set of papers.

1. Symmetry, privacy, and "I"-thoughts: A reply to Valente and Verdejo

In their contributions to this volume (Valente 2026; Verdejo 2026), Matheus Valente and Victor Verdejo offer thoughtful and challenging discussions of a cluster of issues centered on the (putative) shareability of "I"-thoughts. Their papers complement each other nicely. Valente is an opponent of shareability and offers an interesting and original defense of his view, arguing, first, that private "I"-thoughts can be perfectly objective and, second, that "I"-thoughts have to be private if they are to fulfil the role they are widely held to play in explaining and predicting intentional actions. Verdejo, in contrast, is supportive of the general idea behind what I term the Symmetry Constraint (which is that it must be possible for token- utterances of "I" and "you", in an appropriate context, to express the same thought). His paper offers, among other wide-ranging contributions, a new way of accommodating symmetry and shareability.

This reply to Valente and Verdejo begins with a review of Frege's brief but influential discussion of these matters in 'Der Gedanke' ('The thought'), which has set the frame for subsequent discussions. In §1.2, I sketch out the account of "I"-thoughts that I developed in *Understanding 'I': Language and Thought* (Bermúdez, 2017b), building on earlier papers (Bermúdez, 2005a, 2011b), since this account is the target of the two (friendly) discussions. Valente's paper is the topic of §§1.3 and 1.4. In §1.3 I consider Valente's analysis of the (putative) objectivity of private "I"-thoughts, while §1.4 evaluates his ingenious argument from the requirements of psychological explanation to the privacy of "I"-thoughts. The last section, §1.5, considers Verdejo's account of how the Symmetry Constraint is met, which includes fruitful discussions of different levels of analysis for "I"-thoughts and the need to move beyond token-reflexive rules for fixing first-person reference.

1.1. The background in Frege

Frege's discussion of the first-person pronoun "I" in 'Der Gedanke' ('The thought') set the agenda for subsequent discussion. His compact discussion raises, *inter alia*, two important questions. One has to do with understanding. Frege asks, in effect, whether someone else can understand the thought that I express using "I"? This question about understanding is independent of the second question that he raises, which is about thought-identity. Here Frege asks whether someone else can think the same thought that I express using "I".

It is perfectly possible to think that, although nobody else can think the thought that I express using "I", there is no obstacle to someone else understanding that thought. Plainly, if that view is correct, understanding a thought does not require that one think it, or even be capable of thinking it. As will be discussed further below, this is a popular view. It is not clear that it is Frege's view, although his discussion is not terribly easy to follow.

Frege begins as follows:

> Consider the following case. Dr Gustav Lauben says. 'I was wounded', Leo Peter hears this and remarks some days later, 'Dr Gustav Lauben was wounded'. Does this sentence express the same thought as the one Dr Lauben uttered himself?
>
> Suppose that Rudolph Lingens was present when Dr Lauben spoke and now hears what is related by Leo Peter. If the same thought was uttered by Dr Lauben and Leo Peter, then Rudolph Lingens, who is fully master of the language and remembers what Dr Lauben said in his presence, must now know at once from Leo Peter's report that he is speaking of the same thing. (Frege, 1918/1977, p. 11)

The discussion that immediately follows does not really touch upon Dr Lauben's original utterance, focusing instead on the relation between what Leo Peter says and Rudolph Lingens hears. The upshot of that discussion is that, if the two men both think of Dr Lauben in the same way (as, e.g., the only doctor living in a house known to both of them), then they do both understand the thought in the same way. Frege immediately glosses "understanding" as "they associate the same thought with it". Here at least it seems that the questions of thought-identity and understanding collapse into each other. But Frege does not at this point answer his original question about Dr Lauben and Leo Peter.

Frege then modulates into discussing proper names, offering a pithy formulation of his notion of sense in terms of the way in which the object being referred to is *given* or *presented* (die oder das durch ihn Bezeichnete gegeben ist), and expressing the desirability (presumably in an ideal language) of there being only one way of being given/mode of presentation/sense associated with each proper name. There then follow three sentences that have, so to speak, launched a thousand ships:

> Now everyone is presented to himself in a special and primitive way, in which he is presented to no-one else. So, when Dr Lauben has the thought that he was wounded, he will probably be basing it on this primitive way in which he is presented to himself. And only Dr Lauben himself can grasp thoughts specified in this way. (Frege, 1918/1977, pp. 12-3)

This passage is generally taken to commit Frege to the existence of private senses – to the thesis that whenever anyone thinks about themselves they do so using a sense that is unique to them. It is worth pointing out though, some slippage in the passage that has not, to the best of my knowledge, been remarked upon. Frege says that Dr Lauben's "I"- thought (to use a more modern turn of phrase) is *probably based on* (wahrscheinlich [...] zugrunde legen), this primitive way in which he is given to himself. Frege's official view, of course, is that a sense is part of a thought, not something on which a thought is based. And, while he plainly believes that every sense is a way of being given, there are surely ways of being given that are not senses. At no point in this passage (or, to the best of my knowledge, anywhere else) does Frege actually identify "the primitive way in which he is presented to himself" with the sense of "I", or even say that it is a sense at all. It is true that he does speak of a thought that only he can grasp (den nur er allein fassen kann), but this seems too thin to describe as a theory of the sense of the first-person pronoun, or of self-conscious thought.

In any case, Frege acknowledges that private thoughts cannot be communicated and so successful communication requires that "I" be expressed with a sense (Sinn) that others can understand —something along the lines, he suggests, of "he who is speaking to you now". What is typically not remarked upon, however, is that Frege, in a fine example of ring composition, returns to the very question with which he began: "Yet, there is a doubt. Is it all the same thought which first that man expresses and then this one?" (Frege, 1918/1977, p. 13). He leaves the question hanging, and goes on to talk about the general differences between thoughts and ideas.

It seems reasonable to conclude, therefore, that Frege is not here articulating a settled doctrine of how "I" works either in thought or in language.[1] But still, there are some significant take-aways for anyone wanting to tackle first-person thoughts from within a broadly Fregean perspective. The first is the importance of articulating the connection between the sense of a subsentential item such as "I" and what it is for a listener to understand a sentence in which it features. The second is the importance of articulating the connection between understanding a sentence and identifying the person whom the sentence is about, which in this particular case is of course the referent of the indexical "I". And finally, there is the question that runs through Frege's discussion: Does my understanding what someone else says require me to think the very same thought that they are expressing?

In my book *Understanding "I": Language and Thought*, I offered a Fregean account of first-person thought and first-person language that works within these constraints. In §1.2 I will sketch out that account, before turning to the criticisms offered by Matheus Valente and Victor Verdejo, in §1.3 and §1.4 respectively.

1.2. The sense of "I"

Understanding "I": Thought and Language (Bermúdez, 2017b, extending earlier discussions in Bermúdez, 2005a) develops a model of how language-users understand sentences involving the first-person pronoun "I". The model is intended to illuminate the unique psychological role that self-conscious thoughts (typically expressed using "I") play in action and cognition – a unique role often summarized by describing "I" as an essential indexical. As John Perry, David Lewis, and many others have argued, first-person thoughts (thoughts typically expressed through sentences involving "I") have a unique and indispensable role in generating and explaining action. In Perry's famous example, I might curse the shopper spilling sugar as much as I like, but it will not motivate me to look inside my cart until I realize that I myself am the careless shopper (Perry, 1979). Since, sentences using "I" express first-person thoughts, it seems reasonable to think that explaining how we understand the first-person pronoun is one way of explaining what first-person thoughts are and how they work.

Understanding "I" proposes an account of the sense of "I", based on the guiding idea that a theory of sense is a theory of understanding (as opposed, for example, to giving an account of how reference is "fixed"), an interpretation of Frege most closely associated with Michael Dummett (1973, 1981). Frege's notion of sense has been called upon to do many different jobs, but fundamental among these is the job of explaining linguistic understanding – in virtue of what do we understand a sentence, or (equivalently) grasp the thought that it expresses?[2] As outlined in the previous section, Frege himself was very sensitive to the difficulties that arise when thinking about how sentences with indexicals such as "I" are understood.

Following Frege (particularly as read in the Anglo-Saxon tradition inspired by Dummett, Geach, McDowell, and others), the best way to think about understanding is through the notion of truth. We understand a sentence insofar as we grasp its truth-condition, where that latter notion is in turn explained in terms of knowing what it would be to establish its truth-value —in other words, by knowledge of its truth-condition. That latter notion does not have to be understood in the anti-realist manner favored by Dummett, who thinks that one cannot know what it would be to establish a sentence's truth-value unless one can be (at least in principle) in a position to recognize its truth through some sort of effective procedure. Applying this general model to the matter in hand, our understanding of the constituent words of

---

1 For these reasons, therefore, I have become more skeptical than I used to be about attributing to Frege, as Dummett does (1981, p. 123), a distinction between the "I" of soliloquy and the "I" of communication.

2 It should be emphasized that this is a question that any account of meaning, broadly construed, will have to answer. Direct reference theorists and possible worlds semanticists are not immune. And so, the account proposed in *Understanding "I"* was offered in a broadly ecumenical spirit. A broadly Fregean account of linguistic understanding is orthogonal to contemporary debates about the nature of reference or how to understand propositional attitude ascriptions.

a sentence must be based in our understanding of what they contribute to determining the truth-condition of sentences in which they feature. And so, *pari passu*, the sense of the self-referential term "I" is given by our ability to determine the truth-condition of sentences that are appropriately self- referential.

As indicated in §1.1, Frege himself does not go much beyond suggestive remarks, and the best developed neo-Fregean account of the sense of "I" has been offered by Gareth Evans in *The Varieties of Reference* (Evans, 1982). Evans's account is grounded on the basic principle that understanding referring expressions involves standing in, and being able to act directly upon, *information-links* to the referent of the expression. Applying this to the first person, Evans emphasizes information-sources that gives rise to judgments that are *immune to error through misidentification relative to the first-person pronoun*, including:

— introspection (information about occurrent thoughts and mental states)

— somatic proprioception and kinesthesis (information derived from bodily awareness about limb position and bodily movement)

— visual kinesthesis (visually derived information about movement and spatial relations to perceived objects)

— autobiographical memory (information about one's past derived from remembered episodes in one's personal history)

Evans emphasizes two further aspects. In addition to the direct implications of "I"- thoughts for action (as in the thesis of essential indexicality), Evans's self-location component makes mastery of "I" dependent upon subjects' ability to locate themselves in an objective spatio-temporal world, integrating their egocentric path through the world with a cognitive map that is perspective-free.

Evans himself defended the view typically attributed to Frege (but see §1.1 above), on which "I" has a private sense so that thoughts in which it features cannot be shared, arguing (unsuccessfully to my mind) that unshareable "I"-thoughts can nonetheless be objective. In contrast, I argue for what I term the *Symmetry Constraint*: An account of the sense of "I" must allow tokens of "I" to have the same sense as tokens of other personal pronouns such as "you" in appropriate contexts (Note: *must allow*; not *requires*). In certain contexts, not terribly unusual, the very same (token) thought can be expressed by me using "I" and by you using "you". I motivated the Symmetry Constraint in three ways —(a) through considerations of same-saying (as emphasized by Mark Sainsbury in his work on indexicals and reported speech (1998); (b) logically, in terms of when speakers should be counted as contradicting each other; (c) on epistemological grounds, in order to allow paradigm instances of testimony to count as knowledge. The basic argument is that each of these important phenomena has as its core a set of basic cases in which content is shared. These are cases where, for example, (a) I can report what you say in a given context in a way that allows me to use "you" (or even "he") to say the same thing that you said; (b) I use "you" to deny what you are saying using "I", or (c) I can acquire from you directly knowledge of how I am from what you say about me.

Importantly, the Symmetry Constraint is formulated at the level of individual tokens in a specific context. Since, at a more general level, understanding how to use the first-person pronoun "I" is plainly a very different matter from understanding how to use the second person pronoun "you". Accepting the Symmetry Constraint, therefore, requires us to distinguish (at least in the case of "I" and other indexical expressions, such as "now" and "here") between two different types of understanding, and correlatively two different types of sense.

> *The token-sense of "I":* What a speaker or hearer understands when they utter or hear a token utterance involving "I".
>
> *The type-sense of "I":* That in virtue of which a speaker or hearer can properly be said to understand the expression "I".

We need to make a distinction, in other words, between two types of understanding. The Symmetry Constraint applies only in specific contexts (when I, for example, directly addressing you, use a sentence with "you" to deny what you have said using "I"). Those are contexts where, I argue, your understanding of what is said and mine coincide. Whatever account we give of *that* type of understanding, however, it is quite plainly distinct from the understanding of "I" that is required for linguistic mastery of the first-person pronoun. This is the type-sense of "I", which is given by the token-reflexive rule that a token of "I" refers to the producer of that token. Token-sense, however, is more complicated.

The Symmetry Constraint rules out explaining the token-sense of "I" in terms of sensitivity to the types of self-specifying information that Evans emphasizes. My proposal in *Understanding "I"* is that we think about the token-sense of "I" in terms of the third component in Evans's account, namely, the thinker's ability to think of herself as an object uniquely located in space and following a unique trajectory through space- time. Following Evans I think of this knowledge as a practical capacity – the thinker's practical capacity to situate herself within a cognitive map of the world, superimposing her egocentric understanding of space upon a non-egocentric representation of how the environment is spatially configured. This way of thinking about token-sense dovetails nicely with the distinctive functional role of "I"-thoughts, because connecting egocentric and objective space seems to be exactly what happens in the paradigm cases of essential indexicality. Moreover, it allows the Symmetry Constraint to be met, I argue, because those contexts in which "I" and "you" have the same sense are contexts in which you and I can plausibly be described as exercising the same location ability.

* * *

This brief sketch of the Fregean background and the neo-Fregean view developed in *Understanding "I"* will hopefully frame Matheus Valente and Victor Verdejo's rich and stimulating contributions to this special issue. To situate their views at a high level of abstraction, Valente (2026) takes issue with the Symmetry Constraint. He rejects my arguments in support of the shareability of "I"-thoughts, as well as my objections to Evans's claim that unshareable "I"-thoughts can nonetheless be objective, before going on to offer a more general argument based on seemingly uncontroversial theses about the role of propositional attitudes in psychological explanation that he takes to show that "I"-thoughts cannot in principle be shareable. I will discuss Valente's thought-provoking arguments in the next two sections, §1.3 and §1.4. Verdejo (2026), in contrast, accepts a version of the shareability thesis (although not in the form sketched out above), but proposes to explain shareability at the level of reference, rather than at the level of sense. He offers an account of what he terms the self-type, namely, the type of thought expressible with first-person utterances, on which it is possible for thoughts of the relevant type to have as their subject someone distinct from the thinker of that thought. On this view, when I contradict, using "you", what you said using "I", my "you"-utterance is actually expressing a thought of the self-type. This ingenious and original argument will be discussed in the final section, §1.5.

1.3. Symmetry and objectivity: Reply to Valente

Valente's (2026) determined attack on the putative shareability of "I"–thoughts, both in general and in the specific form in which I defend it, has two components. In the first, which will be the topic of this section, Valente tries to undercut one of the motivations for shareability by developing a line of thought that originates with Evans. One very general reason for thinking that "I"-thoughts must be shareable is that this would seem to be required for them to be objective. This is a very natural thought for a Fregean to have, given Frege's insistence on the importance of distinguishing thoughts, as objective, abstract entities, from the contents of individual consciousness. A sense that is necessarily private (the so-called "I" of soliloquy) would, the argument goes, amount to no more than a subjective idea – precisely the sort of thing from which Frege, in 'Der Gedanke' and elsewhere, is at such pains to distinguish genuine thoughts.

Valente agrees with Evans that objectivity does not require shareability. As he (Evans) puts it:

> What is absolutely fundamental to Frege's philosophy of language is that thoughts should be objective – that the existence of a thought should be independent of its being thought by anyone, and hence that thoughts are to be distinguished from ideas or the contents of a particular consciousness. When Frege stresses that thoughts can be grasped by several people, it is usually to emphasize that it is not like an idea.
>
> A true thought was true before it was grasped by anyone. A thought does not have to be owned by anyone. The same thought can be grasped by several people.
>
> His most extended treatment of the nature of thoughts – 'The Thought' – makes it clear that it is the inference from shareability to objectivity which is of paramount importance to Frege, rather than shareability itself. Since an unshareable thought can be perfectly objective – can exist and have a truth value independently of anyone's entertaining it – there is no clash between what Frege says about 'I'– thoughts and this, undeniably central, aspect of his philosophy.[3]

Valente, to his credit, offers an argument to support the attempt to disassociate objectivity from shareability. He begins by considering the following timeless thought:

> (1) MV thinks: (T2) I am tired at 1200 GMT on Sunday January 3, 2025

We are to suppose that MV grasps this thought for the first time at some point in February 2025, although he had for some time been aware that someone was tired at that point in time (and presumably point in space). What MV discovers in February is that he himself is that very person whom he had all along known to be tired at noon on January 3, 2025. Valente writes:

> Intuitively, the truth which I discovered would have remained true, even if neither I nor anyone else had discovered it: after all, what else could it mean for something to be discovered in the first place. But Bermúdez must instead hold that my act of apprehending T2 for the first time is what brings T2 into existence, and so what determines its being a true thought as opposed to nothing at all. This seems to get things the wrong way around. More generally: if first-person thoughts are neither true nor false before apprehension, then in which sense could their truth amount to a discovery? (Valente, 2026, p. 139)

Although talk of discovery in this context might seem somewhat to beg the question, it is hard to disagree with Valente's basic idea that he does discover *something* when he thinks T2 for the first time. But there are many ways of doing justice to that thought without claiming that what he discovers is the content of T2, namely, the thought that it expresses. In fact, it seems odd to talk about discovering *thoughts* at all. Surely, what one discovers in the normal case are *facts*, not thoughts —things about the world, not about how we think about the world. This is not peculiar to token-reflexive or indexical thoughts. When Pythagoras determined that the square of the hypotenuse is the sum of the squares of the other two sides of a right-angled triangle, he *grasped* a thought, but what he *discovered* was a fact about right-angled triangles.

In the case under consideration, what MV discovers is the fact that he was the person who was tired at noon on Sunday, January 3. This is just an ordinary fact, that might equally well be reported with the sentence "MV is tired at 1200 GMT on Sunday, January 3, 2025". "I"-thoughts come into the picture, not as the things that are discovered, but rather as the specific way in which those things are discovered. To return to Perry's famous example, when Perry discovers that he himself is the shopper spilling sugar, what he discovers is that Perry is the shopper spilling sugar. What is interesting about the example, though, is how he discovers it. He discovers it by thinking an "I"-thought that immediately puts

[3] Evans (1981, p. 313), quoting from Frege's posthumous essay 'My logical insights'.

him in a position to act upon the discovery (namely, by looking in his shopping cart for the bag of sugar with a hole in it). There is certainly no need to assume that this requires the "I"-thought to exist prior to being thought, which is fortunate since, as suggested above, the identity of the thought depends upon the act of thinking.

The basic problem for the Evans/Valente argument, then, is that "I"-thoughts cannot exist without a thinker. There is nothing peculiar about "I"-thoughts here. Exactly the same argument goes for any thought that can only be expressed through a token- reflexive, such as "here", "there", "now", "then", "we", "you", and so on. To every token- reflexive thought there corresponds a non-token-reflexive thought, arrived at by replacing any token-reflexive expression with a non-token-reflexive expression. If one were to ask what a fact is, it would be reasonable to respond that a fact is a true thought. But there is not a distinct fact for every true thought. Facts are at the level of reference, not at the level of sense, and so thoughts stand in a many-one relation to facts.

In sum, non-token-reflexive thoughts can unproblematically be taken to exist independently of being thought, precisely because their identity is not fixed by the act of thinking. So too can the facts that correspond to token-reflexive thoughts. That these two things hold is most likely what gives the impression that token-reflexive thoughts can exist independently of being thought. But this impression must be resisted. As I pointed out in *Understanding "I"*, succumbing to it betrays a confusion between two superficially similar conditionals. The first, in which the quantifier has narrow scope, holds that, whenever there is an appropriately formed episode of thinking there is a content for that episode of thinking:

(*) There is an appropriately formed episode of token-reflexive thinking in context <a, t, l> □ → ∃x [x is the content of that episode of thinking]

(*) is trivially true, because an episode of token-reflexive thinking is appropriately formed when and only when it has a content. But changing the scope of the quantifier from narrow to wide transforms a trivial truth into a substantial falsehood. As I have tried to bring out, there are no good reasons to accept, and many good reasons to reject, (**)

(**) ∃x [There is an appropriately formed episode of token-reflexive thinking in context <a, t, l> □ → x is the content of that episode of thinking]

It would seem, therefore, that the prospects are limited for arguing that non-shareable "I"-thoughts can be objective. But perhaps so much the worse for objectivity, if Valente's main argument against shareability is sound. So we should turn now to that argument.

1.4. Valente's Asymmetry Argument against shareability

Valente's contribution to this volume contains a bold and striking argument that, if successful, rules out the possibility of shareable "I"-thoughts. This argument will be the topic of this section. Valente argues that the explanatory power of "I"-thoughts depends upon their being private. His argument rests, of course, on the thesis of essential indexicality — namely that "I"-thoughts, and indexical thoughts more generally, have an indispensable role to play in psychological explanation. This thesis has been challenged (most prominently in Cappelen and Dever, 2013), but since I both accept and have defended it (Bermúdez, 2017b, 2017c) this is common ground between us.

Valente's argument requires a little bit of set-up. He develops the technical notion of a *confidant*. Two individuals are confidants in his sense iff:

(ii) they acknowledge each other as ideally rational agents,
(iii) they have engaged in an idealized communicative exchange whereby they have mutually expressed all their attitudes, with the result that:
(iv) their post-communication attitudes are the result of updating on everything that they have learned from each other.

Confidants in this sense have what Valente describes as *maximally convergent attitudes*. This extends to indexical attitudes, which are *maximally coordinated* in the sense that, if A and B are two confidants, every attitude that A might express with "I" or any other indexical is matched by an attitude of B's expressed with "you", and vice versa (*pari passu* for other indexicals).

It is important to note that indexical attitudes can be maximally coordinated without being identical. That is to say, allowing for the possibility of maximally coordinated indexicals is compatible with both symmetry and privacy. Valente's argument, in brief, is that maximally coordinated indexical attitudes can only play the explanatory role required by essential indexicality if they are private. Shared indexical attitudes would not be able to accommodate the fact that two confidants with maximally coordinated indexical attitudes can behave differently in a given situation. Here is how the Asymmetry Argument goes.

(1) Ideally rational agents cannot have all the same attitudes if they have distinct reasons [TWINS].
(2) Some confidants have distinct reasons.
(3) If Symmetry is true, then confidants have all the same attitudes.

Therefore, Symmetry is false (from 1, 2, and 3)

Certainly, the logic is impeccable, but what about the premises? It is worth having Valente's argument for (2) in view. He offers the following illustration, which is a variant of a much-discussed example of Perry's (also discussed in *Understanding "I"*).

> **Bear Attack**
>
> Ann and Bill are walking in the woods when a bear starts chasing Ann. Ann and Bill both realise that the bear is about to attack Ann (Ann thinks 'I'm about to be attacked by a bear' and Bob thinks 'You're about to be attacked by a bear'), and they both want to save Ann from the attack. They act differently. Ann curls up into a ball and plays dead. Bill, who witnesses the attack from a distance, runs to get help.

It is hard to argue with this. Bob and Ann certainly act differently, and so one might reasonably think that their reasons are different. There is little room for disputing (3), given the assumption that Bob and Ann are confidants. So, the only place to take issue is with the first premise, TWINS.

Fortunately (at least from the perspective of a defender of symmetry), there are good reasons for rejecting TWINS. It is perfectly possible for two people with identical attitudes to have different reasons, because reasons do not, as Valente puts it, supervene on attitudes. The basic reason why was noted some years ago by Aristotle in his discussion of action and what he called the practical syllogism. Aristotle, far ahead of his time on this as on so many other things, did not of course have the concept of a propositional attitude. But he does (in, for example, sections III.9-11 of the *De Anima*) discuss the role of belief and desire. He makes a very plausible case that neither belief nor desire on its own can motivate action. At a minimum, they need to operate in tandem. That claim, however, is a necessity claim, not a sufficiency claim. It is common among contemporary analytic philosophers to hold that we can explain intentional actions by attributing to the agent appropriate beliefs and desires. Hence the expression belief-desire psychology and frequent references to belief-desire explanation. It is to Aristotle's credit, however, that he saw, in his doctrine of the practical syllogism, that actions cannot be explained in terms of beliefs and desires alone.

It is not entirely clear whether Aristotle's notion of the practical syllogism is intended to be a model of how actions come about, a template for action-explanations, or some combination of the two, but however it is understood Aristotle makes clear the ineliminable role of perception in action and action-explanation. Here he is making the point in his own inimitable style.

> Since the one conception (*hupolêpsis*) and proposition (*logos*) is universal, and the other is of what is particular (since one says that a certain sort of man must do such and such a thing, and the other says that this is such and such a thing, and I am that sort of man), this latter belief —not the universal one— initiates motion; or, both do, but this one is more at rest., and the second is not.[4]

Motion for Aristotle comes about from a universal proposition combined with something particular. The universal proposition specifies a given course of action for a particular type of agent in particular circumstance (a certain sort of man must do such and such a thing), but that proposition alone cannot cause anyone to act unless it is "anchored" to the here and now. The agent acts when he realizes that *these* are the right kind of circumstances and that he is an agent of the appropriate type. The *De Anima* passage is slightly confusing, because Aristotle phrases his point in terms of belief. The role of perception emerges much more clearly in *Nicomachean Ethics* VII.3, he is very clear that "anchoring" to the here and now takes place through perception: "The one opinion (*doxa*) is universal, the other is concerned with the particular facts, and here we come to something within the sphere of perception."[5]

There is a simple explanation, therefore, for why Ann and Bill behave differently in the bear attack case, even though they have exactly the same attitudes. Their perceptions are different, which means that their attitudes are differently "anchored" to the here-and- now. Consider how Valente fleshes his example out.

> Suppose, then, that they spotted the bear right after leaving a communication chamber wherein they communicated in the manner proprietary of confidants. Among other things, they talked about how they ought to react in various cases of bear attack, and even considered the very same scenario at which they found themselves later - having whole-heartedly agreed that, were this scenario to become actual, then whoever is under attack would need to curl up into a ball while the other would run for help. (Valente 2026, pp. 145-146)

The explanation of why Bill runs to get help is that he *perceives* that it is Ann who is under attack, whereas Ann curls up into a ball because she sees that the bear is attacking her.

This account of the matter is perfectly consistent with Ann and Bill being maximally convergent and having maximally coordinated attitudes, because perceptions are not propositional attitudes. Whatever one might think about "I"-thoughts and other indexical attitudes, it is plain that perceptions are private and non-shareable. That is why, for Frege, they are paradigms of ideas, as opposed to thoughts. It is equally plain that perceptions provide reasons for action. Combining these two thoughts explains the basic insight that drives Valente's Asymmetry Argument. He is absolutely correct that psychological explanation does depend upon certain private psychological states and, moreover, that those private psychological states explain why different people with the same beliefs and attitudes can act differently in the same situation. He goes wrong, however, in thinking that these private psychological states are indexical thoughts. They are not. They are perceptions.

The only way that I can see for Valente to resurrect the Asymmetry Argument would be to try to bring perceptions back under the scope of TWINS. He could argue either that perceptions are propositional attitudes or, only marginally less implausibly, that psychological explanations do not invoke perceptions, but rather the perceptual beliefs to which they give rise. Neither strategy holds much promise, however. The differences between perceptions and beliefs just seem too overwhelming for the first strategy even to get started. To take just one key marker, perceptions are occurrent states, while beliefs and other propositional attitudes are dispositional states. That alone seems enough to rule out any sort of assimilation argument. It also, with a little argumentation, would seem to rule out the second strategy. It is because perceptions are occurrent that they can inform me about how the world that I perceive is configured, relative to me, at this very moment. And that is the information that is required to anchor intentional action, to allow standing states such as beliefs and desires to be expressed and applied in actions in specific contexts.

---

4 DA III.11 434a16–21, translated in Shields (2016).

5 NE VII.3 1147a25–26, trans. Ross/Urmson in Barnes (1984). For more discussion of Aristotle on action see Ch. 10 of Bermúdez (2025).

Despite the ingenuity of Valente's arguments, therefore, I do not feel compelled to abandon the Symmetry Constraint. The question remains, though, whether my proposed account of how the Symmetry Constraint is met is the best one. Victor Verdejo does not think so, and so I turn to his discussion in the next section.

1.5. Verdejo's account of symmetry

Unlike Valente, Verdejo (2026) is very much in sympathy both with the Symmetry Constraint, as I formulate it, and with my motivations for maintaining it. So, we start from a considerable degree of common ground. It will be helpful for the following to have in front of us an example of the Symmetry Constraint in action on which we can both agree. He proposes the following pair of sentences:

(a) At t_1, X addressing Y, utters: "I am a professional philosopher".
(b) At t_2,Y, addressing X, utters: "You are a professional philosopher".

We are to assume that t_1 occurs shortly before t_2; that (a) and (b) are uttered in a single conversation by two people who are aware of who they are addressing; and so on. In such a situation, we both think, it must at least be possible for (a) and (b) to express the same thought. It might be the case, for example, that, in (b), Y is expressing knowledge acquired through X's utterance in (a). This would be a base case of knowledge by testimony, in which someone comes to know that *p* by being told that *p*. Alternatively, we might slightly adapt the example:

(c) At t_2, Y, addressing X, utters: "You are not a professional philosopher".

Here we might say that, with the same conditions holding, Y in (c) could be denying what X says in (a), understanding denial in a strict and simple sense. This is the sense in which what is said in (c) is the negation of what is said in (a). The logical structure is *p* and ~*p*. It is not *p* and *q*, with an implied premise to the effect of ($q \Leftrightarrow \sim p$).

The challenge, therefore, is to explain what *p* is, and in particular to explain how a single thought can be expressed with the first and second person pronouns. It is plain that (a) and (b) have the same *truth condition*, and that (c) is true just if that truth condition does not hold. But that is not the *explanandum* – at least not in my view. As Verdejo notes, I take the issues raised by the Symmetry Constraint to arise when we focus, not on what makes utterances true, but on what it is to grasp the thought that they express, on the guiding assumption that sentences with the same truth condition can express different thoughts. As we will see further below, Verdejo has doubts about using Frege's own (negative) criterion for individuating thoughts (viz., that two thoughts are distinct just if it is possible rationally to assign them different truth-values), but it is plain that he takes himself to be operating at the level of thought (sense), rather than at the level of truth (reference).

In light of this it is striking that Verdejo offers an account of how symmetry can hold in terms of a rule for fixing the reference of "I". To see why he thinks that specifying a reference-fixing rule can solve a problem at the level of sense we need to look at a helpful distinction he makes between three different levels of considering what he calls the self-type – namely, the type of thought expressible by utterances of the first-person pronoun. These are:

> **The general self-type** = the level at which your utterance of "I am F" and my utterance of "I am F" express the same type of thought.
>
> **The instantiable self-type** = the level at which the general self-type is indexed to a particular individual (e.g., the level at which my utterances of "I am F" at t_1 and "I am G" at t_2 express the same type of thought).
>
> **The instantiated self-type** = what Verdejo terms "unique, unrepeatable episodes of thinking" (Verdejo 2026, p. 124)

Relative to this classification, Verdejo argues that symmetry needs to be explained at the second level, the level of the instantiable self-type. The question he tackles is whether the instantiable self-type expressed by X when he says "I am a professional philosopher" at t_1 can be the same as the instantiable self-type expressed by Y when she says "you are a professional philosopher", at the same time.

Verdejo's account of symmetry is motivated by the basic idea, originally developed in Verdejo (2018) (to which I replied in Bermúdez, 2019a), that shareability can only be accommodated by removing the requirement that the referent of an "I"-thought must be its thinker. I think that we can reconstruct his reasoning as follows. When X utters (a), the thought he expresses is plainly about X. It is also, equally plainly, an "I"-thought. So, if the thought that Y expresses with (b) is to be the same thought, then it too must be an "I"-thought about X. But then Y must be thinking an "I"-thought about X, which would be impossible if "I"–thoughts could only be about their thinker. Since the identity of the thought that X expresses in (a) and the thought that Y expresses with (b) is to be captured at the level of instantiable type, Verdejo draws the conclusion that what is needed to explain symmetry is an account of how the reference of "I" is fixed at the level of instantiable type that lifts the requirement that an "I"-thought must be about its thinker.

Hence, Verdejo proposes the following reference-rule, which he takes to individuate the instantiable self-type:

> **The self rule**: An instance of the instantiable self-type refers to the subject of the thought in which it features.

The guiding idea here is that, while in many cases the subject of the thought will be the thinker of thought, the self rule leaves it open that a given thought might have a subject other than its thinker. So, in the example, just discussed, Verdejo's self-rule allows for Y to think X's "I"-thought. In this case, the thought that Y expresses in (b) has X as its subject. It is, so to speak, X's "I"-thought, which Y happens to be thinking.

It seems clear (to me at least) that Verdejo has accurately described the situation that must hold in cases where the Symmetry Constraint holds. I still do not see, however, how the self-rule as he has stated it can have the explanatory power that he thinks it has. The basic problem is that it rests upon a key notion that is itself unexplained. No account has been given of what it is for something to be the subject of a thought, or of how to determine what the subject of a thought is. Someone who thinks that the reference of an indexical in language or in thought is fixed by some form of token- reflexive rule can given an immediate answer. The subject of the thought is the referent of the indexical, which is fixed by the token-reflexive rule that "I" refers to its utterer/ thinker. But since Verdejo is rejecting the token-reflexive rule, he needs to provide an alternative way of fixing the reference, and the account he proposes cannot do that job on its own.

The difficulty is that (in the absence of any further theoretical articulation of one or the other notion) there is only a notational difference between the statement that X is the subject of the "I"-thought expressed in (b) and the statement that the "I" in the "I"- thought X refers to X. To be the subject of an "I"-thought just is to be the referent of the "I". And for that reason there is no prospect of explaining either notion in terms of the other. The only way of escaping this problem is to provide an independent account of one or the other notion. If one thinks that the referent of "I" is fixed by the token-reflexive rule, then one can use that independent account to explain what it is for something to be the subject of the "I"-thought. But then, *pari passu*, if one wants to explain the referent of "I" as the subject of the "I"-thought then one has to ground that explanatory claim in a plausible account of what it is to be the subject of an "I"-thought.

Verdejo is moving in rather a tight circle when he writes:

> [...] 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. (2026, p. 126)

The problem is that Verdejo is now defining what it is to be the subject of an "I"-thought in terms of the referential condition that is supposed to explain what it is to be the subject of an "I"-thought!

To sum up the discussion so far, I have several points of agreement with Verdejo. First, he is is absolutely correct that doing justice to the Symmetry Constraint requires allowing for the possibility that a first-person thought might be thought by someone other than the person who is thinking that thought. Something like this is surely going on in the three types of cases that motivate the Symmetry Constraint – viz., knowledge transmission through testimony, denial, and same-saying indirect speech reports. Second, it seems reasonable to draw the conclusion, as he does, that allowing for that possibility requires formulating a reference-rule for "I" as it features in thought that is not straightforwardly token-reflexive. Third, Verdejo is surely correct to say that, once the token-reflexive requirement is lifted, it makes sense to appeal to a notion of the subject of a thought in order to articulate to whom the thought refers.

My major concern, however, is that this is the beginning of the discussion, not the end. What Verdejo needs in order to do justice to these insights is a substantive account of what it is for something to be the subject of an "I"-thought. Such an account needs to explain, not just *that* it is possible for the subject of an "I"-thought not to be its thinker, but also *how* it is possible, and *when* it is possible. My own view, very much in line with the account developed in *Understanding "I"* and sketched out in §1.2 above, is that the best (only?) way to do this is through developing a theory of sense *qua* theory of understanding. That is to say, we should approach the question of what the subject of an "I"-thought is through explaining what it is to understand a sentence that expresses the relevant "I"-thought. That in turn will rest upon an explanation of what it is to know that sentence's truth-condition (i.e., knowledge of what it would be to establish its truth- value). I have offered such an account, but of course there may be a better one to be developed.

I end this discussion with some brief remarks on why any such account might not be best formulated at the level of what Verdejo calls the instantiable self-type. To recap, this is the level at which the general self-type is indexed to a particular individual (e.g., the level at which my utterances of "I am F" at *t*1 and "I am G" at *t*2 express the same type of thought). The basic difficulty is straightforward. While it is true that the cases in which we see a distinction between the subject of the thought and the thinker of the thought all occur at the level of the instantiable self-type, the fact remains that the overwhelming majority of "I"-thoughts considered at the level of instantiable self-type will be ones where the subject of the thought coincides with the thinker of the thought. So, if we want to pick out all and only the thoughts where subject and thinker diverge, it would be profitable to operate at a level more fine-grained than the level of instantiable self- type. My own view, as indicated earlier, is that some sort of notion of token-sense is the best place to look. As far as I can see, my notion of token-sense maps fairly closely onto Verdejo's notion of instantiated self-type. However, he construes the instantiated self- type in such a way as to make it impossible for any two instantiated self-types to be the same, because he makes it a matter of definition that instantiated self-types are unique, unrepeatable, and so on. On the face of it, however, this is simply incompatible with the Symmetry Constraint. I cannot see how the thought that I express at *t*2 with "You are a professional philosopher" could be the same as the thought that you expressed at *t*1 with "I am a professional philosopher", without there being a single, repeated instantiated self-type. For that reason, therefore, I think that Verdejo would do well to reject, not the actual distinction between instantiable and instantiated self-types, but the thesis that instantiated self-types are unique and unrepeatable, because that thesis threatens to make it impossible to do justice to the Symmetry Constraint.[6]

* * *

I am very grateful to Matheus Valente and Victor Verdejo for their thoughtful and informative contributions to this special issue, as well as for their careful attention to my own work. As the discussion above shows, I am not fully convinced by their arguments. However, they have certainly put pressure on my views in ways that have helped me personally and also shed significant light on these tricky problems. I look forward to continuing the discussion.

[6] To be sure, there are in the case considered two distinct, unique, and unrepeatable episodes of thinking, but an episode of thinking is not a thought.

2. *'Naturalized sense data' revisited*

I am grateful to Derek Brown both for his thoughtful paper (Brown 2026) and for giving me the opportunity to revisit 'Naturalized sense data', which I find it hard to believe was published 25 years ago. In that paper I argued in support of the following two claims about visual perception:

(a) We perceive ordinary physical objects *directly*, in the sense that we have a direct perceptual access to them that allows us to refer to them demonstratively and that explains why our perceptual beliefs are typically justified.
(b) We perceive ordinary physical objects *mediately*, in the sense that we perceive them in virtue of perceiving parts of their facing surfaces.

Naturalized sense data are (parts of) the facing surfaces of physical objects. We perceive them immediately (because we do not perceive them in virtue of perceiving anything else), and in virtue of perceiving them we directly (but mediately) perceive objects. My claim in 'Naturalized sense data' was that this view, by separating out two notions that are typically run together, ((im)mediate perception and (in)direct perception) has the advantages of both uncritical direct realism and traditional sense datum theory, while avoiding their disadvantages.

Brown is broadly sympathetic to my view, and he is certainly very conscientious in providing the strongest case that he can on my behalf, but he is not convinced. He has two different sets of concerns. The first part of his paper offers a careful and perceptive critique of my argument in support of what I called naturalized sense data, which he thinks moves too quickly. The second part of the paper proposes an intriguing objection to my view, which he thinks cannot handle the problem cases of delusion and illusion that he takes to be the strongest motivation for adopting the traditional version of the sense datum theorem. In particular, he thinks that (unlike the traditional version) my account cannot deal with what he calls the Problem of Perception, which is the problem of (a) explaining the differences between veridical, illusory and hallucinatory perceptual experiences, while (b) explaining the phenomenal similarities that can occur between them.

Brown makes a number of telling points, but I feel that the naturalized sense datum theorist is not without resources to reply to them. In the following I will confine myself to two issues, one from each part of his discussion. § 2.1 considers Brown's claim that, despite my argument to the contrary, a traditional sense datum theorist can indeed use the concept of deferred ostension to explain how reference to ordinary physical objects is achieved, even though only (private) sense data are immediately perceived. § 2.2 sketches out a strategy with which the naturalized sense data theorist can tackle the Problem of Perception.

2.1. Sense data and deferred ostension

Here is a brief recap of the argument to which Brown is responding. The starting-point is the following constraint:

> **The Reference Constraint**:
>
> If it is indeed the case that we make demonstrative reference to three-dimensional material objects, then our account of the immediate object of perception should explain how this is possible.

In brief, an account of the immediate objects of perception must earn its keep, so to speak, by explaining how we are perceptually in contact with ordinary material objects, and, so I claimed, this is something that the naturalized sense data theory can do, but not the traditional sense datum theory. My idea was that each theory will have to offer an explanation that takes the following form:

> A given individual is able to refer demonstratively to an object O because he immediately perceives something that stands in relation R to O.

For the naturalized sense datum theorist, relation R is the mereological part-whole relation and demonstrative reference is an instance of the well-studied linguistic device of deferred ostension, whereby one refers to something through ostending some part of that object. Plainly, traditional sense datum theorists cannot appeal to the part-whole relation. In fact, their only plausible candidate for representation R is the representation relation. Sense data, as traditionally construed, cannot be parts of ordinary physical objects, but they do represent them. How then is the reference constraint to be satisfied? In particular, how does ostension work, when the immediate object of perception is a traditional sense datum?

Sense datum theorists have two options, I argued. On the one hand, they can appeal to some form of deferred ostension. Alternatively, they can maintain that ostension by-passes the sense datum and "latches" directly onto the material object. Neither option can be made to work, I argued, leaving the naturalized sense datum theory as the only viable candidate.

My argument for the first horn of the dilemma was certainly too quick, as Brown points out. I wrote:

> Any comprehensive account of deferred ostension will be formidably complex (Nunberg 1978). For present purposes, though, the important point is that (if the demonstrative pronoun latches onto the sense datum, rather than the material object) ostension could only be deferred across the representation/causal relation if the speaker intended and the hearer grasped a suitable principle governing the deferral of ostension from sense datum to the object which causes it and which it represents. It seems obvious, however, that no such principle is implicated in everyday demonstrative reference to material objects. (Bermúdez, 2001, p. 372)

This latter claim Brown takes to be true, but irrelevant. He takes demonstrative reference to be an act performed by a single thinker —a mental act, in essence. And he thinks that it is perfectly possible for this mental act to be performed without the thinker exploiting any sort of general principle connecting sense data to what they represent:

> Suppose that I am alone and possess adequate cognitive and perceptual capacities (setting aside the practical questions of how these things came to be). Suppose also that I correctly believe that SDT is true. Currently, I immediately perceive a blue, oval SD_m and believe that it is caused by something in the physical world. I suggest that I could internally demonstrate the SD_m with the intention to refer to what I believe is the distal physical cause of the SD_m. Such an internal demonstration might be accomplished via a focused attention on the SD_m, and this could be combined with the cognitive intention to refer to what I believe is the distal physical cause of the SD_m. In fact, the SD_m is caused by a robin egg. Thus, the egg is the relevant cause of the SD_m, the SD_m resembles the egg in relevant ways (i.e. both are blue and oval), and my intentions are straightforward. I see no reason why my efforts to generate a deferred reference to the egg would fail. (Brown 2026, p. 166)

Many philosophers would want to take issue with the very idea of a private ostensive act of this type. Someone of a Wittgensteinian persuasion (and I have leanings in that direction myself, at least as far as skepticism about private languages is concerned) would object to the *privacy* – after all, traditional sense data, if such there be, are not just private, but necessarily private. But one might also wonder about how the actual ostension is supposed to work. There is something rather puzzling about the idea that, when we are presented visually with a blue oval object, we are able to distinguish the sense datum that is the immediate object of perception from the blue oval object that seems to be there in such a way that we can (mentally) point to it (the sense datum) and refer to it in thought. One way of putting the plausible intuition behind transparency theories of perceptual experience is that there does not actually *seem* to be anything besides the blue oval object. As a datum about phenomenology, this is hardly an objection to the sense datum theory, which is a theory about how perception actually functions. But it does point to a difficulty with the idea that one might ostend a sense datum. Ostension is at bottom a selective, or contrastive, notion. When I ostend something, I single it from among other things, or against a background. It is not hard to see how I might single out a blue oval object from

other objects, or against a background. But what is the contrast class against which a blue oval sense datum is selected. Is it the blue oval object? Other objects? Other sense data? It is not obvious how to answer these questions.

Let us set these worries aside for the sake of argument, though, to meet Brown on his own ground. The main problem with his proposal is that it makes demonstrative reference a hostage to theoretical mastery. His proposal entails that the mental act of private demonstrative reference is only available to thinkers who are convinced of the truth of traditional sense datum theory. It is not enough for me to ostend the private sense datum. I must ostend it *qua* private sense datum and do so with the intention of referring to its distal cause. To see why this is a problem, recall that we are trying to show how the reference constraint is satisfied. We are trying to show how we succeed in referring to three-dimensional material objects, assuming that we do so succeed. The 'we' here is very inclusive. In fact, it extends exactly as far as the practice of referring to three-dimension material objects. We need an account that will work for everyone and everything that engages in the practice of reference. It is simply not credible that all who engage in the practice of referring to material objects do so through forming intentions to refer to the distal causes of privately ostended sense data. So, even if it is granted that one can privately ostend sense data, such private ostension is not going to help show how the reference constraint is satisfied.

Brown inadvertently illustrates this point with the complexity of his (admittedly ingenious) account of how an interpersonal linguistic practice of demonstrative reference might emerge:

> [...] assume B is standing next to A and the egg is in front of A. This means that B is experiencing private SD that represent A and A's location relative to B, and vice versa. When B's gaze is directed in front of where A is represented to be, B also experiences a blue, oval SD_m (or least what B would describe as a blue, oval SD_m). In this way, B and A could come to single out the same object – the egg – as the referent of A's deferred demonstrative [...]. Consider, finally, a more evolved stage of communication where, as a community, deferred demonstration of this sort is the presumed interpretation of demonstrative expressions of this sort. Thus, A simply says "That is my favourite egg" and B correctly understands A to be demonstrating her blue, oval SD_m and intending, with 'that', to refer to the hypothesized egg. (Brown 2026, p. 167)

Perhaps Brown is right and this type of linguistic practice could emerge in a community of highly reflective and self-conscious sense datum theorists. But it is surely not how we do things now. The naturalized sense datum theorist, in contrast, can give a much simpler account, which starts with my pointing to the facing surface of the egg. In my linguistic community, and every other I have come across, pointing to the facing surface of the egg is normally understood as pointing to the entire egg, including the parts that I cannot see - if I want to be understood as pointing just to the facing surface then I normally need to make that explicit. In virtue of that generally understood principle, I refer to the egg and I am taken to be so referring. Not very complicated and not very mysterious.

My sense is that Brown and I are trying to answer different questions. I just want a simple explanation of how we succeed (all the time and pretty much without fail) in referring to three-dimensional material objects that is not just compatible with, but is actually underwritten by, that fact that we perceive immediately only the facing surfaces of material objects. I think, though, that Brown is trying to respond to some version of the veil of perception objection to traditional sense data theories. He is trying to show how, even if we suppose that a perceiver has access only to private sense data, reference to three-dimensional objects could still be achieved (thereby allowing the perceiver to escape the threatened veil of perception). But, as I have tried to bring out, an answer to the second question does not give an answer to the first.

2.2. The Problem of Perception

Brown's principal reason for preferring traditional sense datum theory to the naturalized sense datum theory is that only the former can, he thinks, provide a satisfactory answer to what he terms the Problem of Perception. This is the

problem of (a) explaining the differences between veridical, illusory and hallucinatory perceptual experiences, while (b) explaining the phenomenal similarities that can occur between them. He points out, quite reasonably, that the naturalized sense datum theory, as I formulated it in the 2000 paper, has nothing to say about cases of hallucination or illusion, and seems on the face of it to be not obviously applicable to them. In the case of hallucinations, for example, one might think that there is no facing surface, and hence no immediate object of perception, while in cases of illusion, although there might be a facing surface, there is little fit between it and the content of perception. In contrast, he suggests, traditional sense datum theory has a neat and satisfying solution:

> Veridical, illusory and hallucinatory experiences are "latching" onto the physical world to different degrees, and in this regard these experiences are different types of perceptual states. Nonetheless, while we can at times subjectively distinguish between veridical, illusory and hallucinatory experiences, at times we cannot. In principle, these different types of experiences can be subjectively indistinguishable. For many, the most powerful justification for SDT is its ability to explain this conundrum. SDT postulates a type of mind-dependent object – SD – as a perceptual intermediary between the perceiver and the physical world. This intermediary is present regardless of whether one's experience is veridical, illusory or hallucinatory – SDT is a "common factor" theory of perception. (Brown 2026, p. 168)

I am not sure that this is as good an explanation as Brown thinks it is, but in this section I will simply sketch out how the naturalized sense datum theorist might tackle cases of illusion and hallucination.

The first point to make is that the naturalized sense datum theory is not in itself a theory of perceptual content. It is an account of where the perceived-in-virtue-of relation bottoms out, so to speak. If we perceive something immediately then we do not perceive it in virtue of perceiving anything else. The very way in which the perceived-in-virtue-of relation is formulated makes clear that the paradigm case is one where the chain of iterations starts with an ordinary material object. The claim is that, when we do perceive an ordinary material object, we do so in virtue of perceiving part of its facing surface. If this is true, then it is true irrespective of how the ordinary material object is perceived or misperceived. If, through some trick of the light, I happen to misperceive an apple as an orange then I must be perceiving the apple. The naturalized sense datum theory says that when this happens I am perceiving the apple in virtue of perceiving part of its facing surface. It would say exactly the same thing in the non-illusory case where I veridically perceive the apple as an apple. In this sense at least, naturalized sense data function as "common factors" across veridical and illusory perception.[7] But they are not common factors with respect to content. So, what explains the difference between the veridical case and the illusory one? Let's focus on Brown's own example:

> Suppose you look at a purple cylindrical bottle and have an experience of that bottle as a blue cylindrical bottle. That is, the blue that you experience isn't experienced as a feature of just any cylindrical bottle, it is experienced as a feature of the very physical cylindrical bottle that sits before you (a bottle which is in fact purple). This is what makes this an illusion: you are actually experiencing the object in question (i.e. there is successful object perception), but you are experiencing it to have a feature that it doesn't have (i.e. there is an error in property perception). (Brown 2026, p. 169)

So far so good. However, Brown's discussion of this example makes clear that he understands the error in property perception in a very particular way. He thinks that this type of illusory perception involves perceiving a property that is not instantiated in the perceived object —the property of blueness. Once this is accepted, then it seems reasonable to ask where the perceived property of blueness is, and then the traditional sense datum theory starts to look attractive. The blueness is not a property of the bottle, which is purple, but it is a property of the sense datum.

---

[7] Brown acknowledges this on pp. 170-171. He does, however, think that I hold that the common factor extends to hallucination. As the discussion below will show, I think that hallucination needs to be treated separately.

This assumption should be resisted, however. The argument only gets off the ground if we accept that in this case there is an experienced property of blueness. But that should not be accepted. We should not move from

My experience has a property F

to

There is a property F that I am experiencing.

Consider somatic sensations such as hunger, and assume that there is a fact of the matter as to whether or not one is hungry such that one can feel hungry without being hungry. Now consider a scenario where this latter case actually holds. I have an illusory feeling of hunger (perhaps because I have taken a pill with exactly the opposite effects of Ozempic). It would be perfectly admissible philosophically, although somewhat labored from a stylistic point of view, to say that my experience has the property of hunger. I feel no pressure, however, to objectify this property by saying that this is some hunger than I am experiencing. A much more natural way of proceeding would be to say that Inverse-Ozempic has caused me to have an experience very similar to that which I have when I am hungry. There is no need to look for some illusory hunger that I might be experiencing. The same holds for the visual case. There is a sense in which, in the scenario Brown is considering, my experience has the property of blueness. But that is only to say that I am having an experience very similar to that which I have when I am experiencing a blue object. There is no need to look for an experienced property of blueness, and hence no need to posit a sense datum to be the bearer of that experienced property of blueness.

Brown and those who think like him are likely to object that this maneuver does nothing to explain why my experience is the way it is. Why am I having a blue-type experience when what is in front of me is a purple bottle? The traditional sense datum theorist has (he thinks) a satisfying explanation - namely, I am perceiving a blue sense datum, but my proposal leaves the blueness of my experience mysterious.

This is hardly compelling, however. Let's go back to the hunger case. I can provide a perfectly satisfying explanation of how Inverse-Ozempic could produce in me an experience of hunger. It could work, for example, by stimulating production of the ghrelin hormone, which then causes contractions in the stomach - or something along those lines. Surely a similar, broadly speaking physiological, explanation can be given for illusory experiences of blue. I don't need to be experiencing something that is blue. I just need to be in a situation where the mechanisms that are normally triggered by (the facing surfaces of) blue things happen to be triggered by a purple bottle. In fact, this is a much more satisfying explanation than appealing to perception of a blue sense datum. Sense datum theorists still need to explain why, in this particular case, there is a blue sense datum rather than a purple one. If they appeal to a similar physiological explanation, then it is hard to see what work the sense datum is doing. It just seems to be an idly spinning wheel. But I have no idea what a non-physiological explanation would look like.

In cases of illusion, then, there is some story to be told about how, through some sort of deviant causal chain that starts with the facing surface of the object being perceived, a non-standard experience is produced. This sort of approach will plainly work for some cases of hallucination, because the line between illusion and hallucination is a little fuzzy. If severe sleep deprivation causes me to see a stationary bush as a hooded man coming towards me, then that seems to be a hallucination. But there seems no reason in principle why this could not be explained in a similar manner to the purple bottle that looks blue – the sleep deprivation and the light, let us say, jointly create a deviant causal chain that starts with the facing surface of the bush and leads me to have an experience similar to that which I would have were a hooded man to be coming towards me. Some fine-tuning would be required for cases where what is hallucinated is an imaginary object – a troll, for example. But this is surely no more mysterious than the fact that I might dream that a troll is coming towards me. Unless sense datum theorists are prepared to argue that dreams are really perceptions of sense data, then they ought to accept that there are, at least in principle, perfectly good explanations available of what might be thought

of as sleeping hallucinations. But then, surely those mechanisms can be brought into play to explain those cases of hallucination that seem most distant from cases of illusion —Macbeth's dagger, for example. Perhaps there will always be some sort of facing surface in play. But perhaps not. Either way, it seems plausible (to me at least) that some sort of deviant causal chain account will be available.

There is of course much more that could be said to build on these tentative remarks. But hopefully enough has been said to make clear that what Brown calls the Problem of Perception is unlikely to provide a decisive reason to prefer traditional sense data to naturalized sense data.

3. Elusiveness and the phenomenology of the self: Response to Gallagher and Serrahima

Shaun Gallagher (this volume) and Carlota Serrahima (this volume) raise interesting and important questions about the nature of bodily awareness and the role(s) of the body in exteroceptive perception. Their approaches are very different. Gallagher comes at the issue through a critique of the idea, found in Husserl and many of his successors within a broadly phenomenological tradition, that the body functions as a sort of zero-point for how the subject experiences the world and acts within it. Serrahima, in contrast, focuses on the relation between what I have called ψ-ownership and ϕ-ownership - namely, and to put it as neutrally as possible, the ways in which we are conscious of ourselves psychologically and physically. These differences in starting-point notwithstanding, the two papers highlight different aspects of a problem that is at the heart of how we need to think about the phenomenology of embodiment. This is the problem of reconciling the phenomenology of embodiment with what is often called the elusiveness of the self, because the most obvious way of taking the phenomenology of embodiment is as a way of being directly presented with the embodied self. In one form or another, the problem of reconciling the elusiveness thesis with the phenomenology of embodiment is something that I have been wrestling with for a long time. I welcome the opportunity to revisit it.

The first section of this comment focuses on Gallagher's discussion of the concept of the zero-point. Steering clear of the finer points of how Husserl, Merleau-Ponty, and their followers should be interpreted, it nonetheless seems to me that the passages that Gallagher presents can be read in a way that avoids the objections that he makes. He makes sound and effective points about the phenomenology of embodiment, and argues, surely correctly, that we do not experience the body as a zero-point. But, on at least one way of interpreting the reasoning that lies behind the concept of a zero-point, the phenomena that he highlights are perfectly consistent with the claim that the perspective from which we experience the world and act within it does indeed behave in the manner of a zero-point. But still, there is a puzzle. The zero-point is systematically elusive, and so one can ask: how can our experience of the world have its origin both in an elusive zero-point and in what Gallagher correctly describes as "thick and articulated bodily schema" (Gallagher, 2026).

As emerges in section 3.2, that same tension between elusiveness and the phenomenology of the body is central to Serrahima's paper. She focuses on my treatment of ψ-ownership and ϕ-ownership. While she has considerable sympathy for my overall approach, she thinks that I end up not doing full justice to the phenomenology of bodily awareness, because I overemphasize the elusiveness of the self. She holds, as I do, that we should aim to give parallel accounts of ψ-ownership and ϕ-ownership, but, because she thinks that there is a fundamental tension between the elusiveness thesis and taking the phenomenology of ownership at face value, and she finds the phenomenology of bodily awareness compelling, she ends up wanting to downplay elusiveness.

Central to both papers, therefore, is the idea that it is problematic both to hold some form of the elusiveness thesis and to think of bodily awareness as a form of self-consciousness. The final section of the paper tries to reconcile this tension between elusiveness and bodily phenomenology. Revisiting Wittgenstein's influential distinction between uses of "I" as subject and uses of "I"-as-object and using it to distinguish notions of the "I"-as-subject and "I"-as-object, I argue that elusiveness should be viewed, not as a deep fact about the self, but rather as a byproduct of the irreflexivity of consciousness —of the fact that mental acts cannot be their own objects. Our awareness of our own bodies through somatosen-

sory awareness and exteroceptive perception is *both* elusive *and* a direct presentation of the "I"-as-subject. The paper ends with some tentative remarks suggesting that introspection may have a comparable structure.

3.1. Gallagher on the zero-point

The concept of the zero-point is both evocative and hard to pin down. Shaun Gallagher has done a valuable service in bringing together a wide range of different discussions, from both the phenomenological and analytic traditions, identifying both commonalities and differences. Most of these discussions have their origin in the writings of Husserl. Here are two suggestive passages with which Gallagher begins his paper:

> [A]ll spatial being necessarily appears in such a way that it appears either nearer or farther, above or below, right or left [...] The lived body then has [...] the unique distinction of bearing in itself the zero point (Nullpunkt) of all these orientations. (Husserl, 1989, §41a; also see Husserl, 2006, §5)
>
> The lived body is, in the first place, the medium of all perception[...]. Obviously connected with this is the distinction the body acquires as the bearer of the zero point of orientation, the bearer of the here and the now, out of which the pure Ego intuits space and the whole world of the senses. (Husserl, 1989, §18)

As Gallagher points out, connecting Husserl's remarks to Gareth Evans's important discussion in *The Varieties of Reference* (Evans, 1982), the key idea is that action and perception take place relative to an egocentric frame of reference. Other passages that Gallagher cites from Husserl and from his successors develop and embellish this basic idea in different ways, but the core notion running through all the discussions is (as best I can make out) this idea of an egocentric frame of reference.

That core notion raises a number or important questions, including the following:

a) How many frames of reference are there for action and perception? Surely more than one! But does that mean that there is one zero-point, or multiple?
b) What is the geometry of these frames of reference? Do they all have a Cartesian geometry?
c) How do the multiple frames of reference coordinate? Is there a single overarching frame of reference that "anchors" the others?

In keeping with the phenomenological focus of Gallagher's paper (and Serrahima's, for that matter), I want to focus instead on what it means, from the point of view of first-person phenomenology, for the body to be "the bearer of the zero point of orientation" or "bearing in itself the zero point of all these orientations."

Gallagher raises the phenomenological issues here in the following manner:

> Moreover, with respect to the perceptual field, the phenomenology of vision is more complex than what's in focus or in focusing. It is never just perception of a focal point, but always includes a periphery, a horizon – a visual landscape arrayed in front of us.[8] The important question is whether that visual landscape refers us back to something of the lived body that we experience prereflectively as we experience the world. Even if a positive answer recommends itself, is that experience ever an experience of a zero-point? (Gallagher, 2026, p. 184)

[8] Things are complicated in vision, not only physiologically, but as one gets closer to the conscious aspects of vision. Aspects of vision that we rarely notice but that we can be made aware of include binocular disparities (required for depth perception), motion parallax, and physiological diplopia. In the case of the latter, for example, an object outside the point of focus may be seen as double. This is typically edited out/suppressed by the brain so we are unaware of it, but some children notice it and complain of double vision. In fact, they have normal vision and they simply have to learn to ignore the diplopia.

Gallegher gives a number of reasons for answering his final question (whether experience is ever experience of a zero-point) in the negative, including references to my own work on the spatiality of bodily awareness, and offers what he takes to be a competing description of what he terms the *agentive situation*:

> The situation of an agent is not equivalent to the external environment; rather, the situation includes the agent and environment or the agent-in-environment. It's defined in relation to the performance of actions structured by varying affordances and characterized by different time scales and degrees of and kinds of intentionality (Pacherie, 2008). On this conception we should not think of the origin of the action as a zero-point; rather, experience originates in a relational system, the structure of which includes elements of the environment (and in this case social/ intersubjective elements), and where a valanced world of affordances is already pushing and pulling a "thick and deep" embodied subject/agent characterized generally by affectivity, its own past history and future interests, and specifically by its ongoing material engagement, all of which figure into structuring perception and action. (Gallagher, 2026, p. 191)

I do not want to take issue with any of this. Gallagher is surely correct to maintain that the body is not experienced *as* the zero-point of orientation. But it is not clear (at least from the passages cited) that Husserl and other proponents of the concept of a zero-point actually think that the body is experienced as a zero-point – i.e., that our prereflective experience of the body is experience *of* a zero-point. One of the passages from Husserl that Gallagher himself cites seems to suggest that the zero-point cannot itself be experienced:

> The body then has [...] the unique distinction of bearing in itself the zero point of all these orientations. One of its spatial points, **even if not an actually seen one**, is always characterized in the mode of the ultimate central here: that is, a here which has no other here outside of itself, in relation to which it would be a 'there'. (Husserl, 1989, §41, my emphasis)

It is hard to read the bolded passage as saying anything other than that the zero-point is not itself part of the content of perception in any explicit sense.[9]

That does not mean that there is no phenomenological difficulty here. Quite the contrary! We just need to make sure that we are asking the right question. That question is not whether prereflective experience of the body is ever an explicit experience *of* a zero-point, but rather what it is like for experience to be *from* a zero-point. That is to say, what is it like to experience the world from the origin of one or more egocentric frames of reference?[10]

In thinking about this question, I think we need to start from precisely the point just highlighted in the passage from Husserl, namely, that the origin of the frame(s) of reference, the zero-point, is not itself a direct object of experience. There is a structural similarity between the thesis of the zero point and what is often termed the elusiveness thesis, which arguably originates with David Hume, but whose more recent exponents include Jean-Paul Sartre and Sydney Shoemaker. According to the elusiveness thesis, which is prominent in Serrahima's paper, the self cannot be the *object* of introspective awareness. The self is the subject of introspective awareness, but cannot be its object.

Whereas Hume's Elusiveness Thesis is directed at introspection, the basic idea behind the zero-point thesis is that perception, and conscious experience more generally, is always from a perspective that it is not itself part of the content of what is. We perceive the world relative to an egocentric frame of reference – more accurately: a range of egocentric frames of reference). But the origin of those frames is not, and cannot be, part of the content of perception. That, I believe, is where the 'zero' in 'zero-point' comes from.

---

9 The qualification is necessary because, as Gallagher has pointed out to me, Husserl may well have held that we experience the zero-point in an implicit and prereflective sense (which is, in fact, Gallagher's own view).

10 Answering this question would be one way of making precise the idea, canvassed in the previous footnote, that the zero-point is experienced implicitly.

3.2. Serrahima on explaining ownership

Broadly speaking, when philosophers talk about ownership they are talking about what it is to experience one's conscious physical and psychological states as one's own. As Serrahima brings out in the beginning of her paper, discussion in this area often focuses on whether this experience is associated with, or grounded in, a phenomenological "feeling" or "sense" of ownership. She thinks, as I do, that this is not the most promising direction to pursue, and follows my suggestion that we should focus instead on self-conscious judgments of ownership —e.g., self-conscious judgments of the form 'I am sad', 'my foot hurts', or 'I am looking at the stadium', and consider both the descriptive/empirical question of how and why they are made, and the normative question of what *reasons* or *warrant* we have for making them.[11]

Serrahima's focus is on how we answer these questions for two different types of first-person judgments. The first are psychological self-ascriptions, made on the basis of introspection – e.g., "I am thinking about the self" or "I am feeling down". These are judgments reflecting psychological ownership (ψ-ownership). The second group are bodily self-ascriptions, made on the basis of proprioception and somatosensation – e.g., 'I have a headache' or 'I am hungry'. These reflect physical ownership (ϕ-ownership). Serrahima thinks, as I do, that the ideal in this area would be an *integrative* account of ownership, containing parallel, structurally similar, and interconnected accounts of ψ-ownership and ϕ-ownership.

Serrahima takes me to task for (as she sees it) following my own prescription in the wrong way, focusing in particular on my 2003 paper 'The elusiveness thesis, immunity to error through misidentification, and privileged access'.[12] In particular, she feels that I place too much weight on the elusiveness thesis in thinking about ψ-ownership and that this leads me to a distorted view of ϕ-ownership, one that neglects the phenomenology of bodily awareness. It will be helpful for the following to retrace my reasoning.

A principal aim of the account of my ψ-ownership was to capture the fact that psychological self-ascriptions are, to use Sydney Shoemaker's phrase, identification-free. When I judge, on the basis of introspection, that I am ψ-ing, for some psychological predicate ψ, this is not based on any identification of myself as the person who is ψ-ing. I am introspectively aware of ψ-ing, but I do not need to identify myself as the person who is ψ-ing. The identification-free nature of psychological self-ascriptions is coeval with the status of psychological self-ascriptions as being immune to error through misidentification (IEM) relative to the first-person pronoun. I cannot, if my psychological self-ascription is based upon introspection, correctly and with warrant judge that someone is ψ-ing but mistakenly identify myself as the person ψ-ing.

That psychological self-ascriptions are identification-free and have the IEM property seems indisputable (although there may be borderline cases with certain forms of delusions). Nonetheless, we still need to explain how, and with what warrant, someone might move from introspective awareness of ψ-ing to the self-ascriptive judgment "I am ψ-ing". Because there is no direct introspection of the "I", subjects making self-ascriptions cannot be taking their experiences at face value. So, how do they arrive at introspective self-ascriptions?

In my initial attempt to offer an explanation, I leant heavily upon Christopher Peacocke's thesis that there is an *a priori* link between being introspectively aware of ψ-ing and being oneself the person who is ψ-ing.[13] My proposal was that what allows subjects both to make self-ascriptive judgments and to be warranted in so doing is their mastery of what I termed a simple theory of introspection (explicitly by analogy with the simple theory of perception and action canvassed in Evans (1982) and Campbell (1994)). As I saw it at the time, mastery of this simple theory of introspection just consists in grasping the *a priori* link between being introspectively aware of ψ-ing and being oneself the person who

[11] For more on the proposed methodology and skepticism about the sense of ownership, see Bermúdez (2011a, 2015, and 2017a), the last two of which are reprinted in Bermúdez (2018e).

[12] Reprinted in Bermúdez (2018e).

[13] For Peacocke's discussion see Ch. 6b of Peacocke (1999).

is ψ-ing. In other words, I move from introspective awareness of ψ-ing to the self-ascriptive judgment "I am ψ-ing" because I grasp that it is *a priori* that I can only be introspectively aware of ψ-ing if it is indeed me who is ψ-ing.

With that account of psychological self-ascription in hand, and in conformity with the principle that parallel accounts should be given of psychological and bodily self-ascriptions, I proposed a simple theory of bodily awareness. There can be no *a priori* link between aware of ϕ-ing through proprioception and other forms of bodily awareness and it being the case that one is oneself the person who is ϕ-ing, because it seems perfectly possible to be hooked up to someone else's body so that one experiences sensations from their body and, unlike in the introspective case, there is no immediately compelling argument to say that this would make them my sensations. But still, it seems reasonable, given the way things are, to say that there is a *de facto* link between being aware of ϕ-ing through bodily awareness and being oneself the person who is ϕ-ing. This opens up room for a simple theory of bodily self-ascriptions, on which bodily self-ascriptions are made and justified in virtue of mastery of the *de facto* link between being aware of ϕ-ing and being oneself the person who is ϕ-ing.

Serrahima finds this strategy for providing an integrated account of ψ-ownership and ϕ-ownership unconvincing. The problem comes with somatic self-ascriptions, because she thinks that it is a mistake to offer an account of ϕ-ownership that does not involve the phenomenology of bodily awareness —a mistake that is exacerbated in my case because I have an independently motivated account of that phenomenology that I could have used. She writes:

> What is then surprising is that, once we have this picture of the content and phenomenology of bodily sensations, which Bermúdez motivates independently, we have the tools available for more than just a "more complicated," yet "simple" theory of somatosensation. I have argued above that, if there is a phenomenology of bodily ownership, then this constitutes a very good explanation of how subjects make judgments of ϕ-ownership compellingly, as well as a prima facie good explanation of bodily IEM. Bermúdez admits that there is such phenomenology. But if there is such phenomenology, and we confer to it its due explanatory and causal role, then the mechanism involved in judgment formation is unambiguous: the subject just takes her bodily experiences at face value given their phenomenology. This is a kind of theory quite different from the simple theory of introspection —and a better kind of theory for that matter. (Serrahima, 2026, p. 209)

Her diagnosis of what has gone wrong is that I am giving too much weight to the elusiveness thesis. The elusiveness thesis seems to entail that there can be no role for phenomenology in an account of ψ-ownership —hence the appeal of my simple theory. But then, if we want an integrated account of the two types of ownership, it is hard to see how we can incorporate bodily phenomenology in an account of ϕ-ownership, which then points to an analogous simple theory in the case of ψ-ownership.

I have to admit that I no longer find the simple theory of bodily self-ascription compelling, at least not in the way that I was trying to use it. That we do exploit some form of simple theory in making bodily self-ascriptions on the basis of somatosensation and other forms of bodily awareness seems very plausible, but Serrahima is surely right that this cannot be all that there is to ϕ-ownership. She herself is attracted to an integrated account that travels in the opposite direction, one that takes phenomenology seriously for ψ-ownership, and hence downplays and/or denies the role of elusiveness in ϕ-ownership, but she does not go into any details in this paper. Since I find the elusiveness thesis compelling, I would prefer not to follow her down that road. But since I would prefer to stick with an integrated account, that leaves me with the task of reconciling elusiveness and bodily phenomenology, which I will try to do in the next section.

3.3. Reconciling elusiveness and bodily phenomenology

The two previous sections will hopefully have brought out an interesting and important problematic emerging from both Gallagher and Serrahima's papers. I have suggested that one important strand in the notion of a zero-point, origi-

nating in Husserl, is coeval to the idea of the elusiveness of the self, originating (perhaps) in David Hume. Gallagher offers convincing reasons for thinking that the self is not experienced *as* a zero-point (that there is no experience *of* the zero-point), but if my interpretation is correct, that is actually built into the notion. The zero-point is supposed to be elusive. What we need to explain is what it is to experience the world from an elusive zero-point that is not itself experienced. At the same time, though, as both Gallagher and Serrahima point out, we seem to be directly presented with our embodied selves, both in somatosensory experience and in ordinary exteroceptive perception. That there is a tension between elusiveness and what certainly seems to be direct experience of the embodied self is explicit in Serrahima's paper and also runs through Gallagher's paper. In this final section, I will sketch out a view on which these two aspects of first-person phenomenology might be able to co-exist.

Let us take a step back to re-examine an important passage from Wittgenstein's *Blue Book* that has, for better or worse, fueled much of the debate in this area.

> There are two different cases in the use of the word 'I' (or 'my') which I might call 'the use as object' and 'the use as subject'. Examples of the first kind of use are these: 'My arm is broken', 'I have grown six inches', 'I have a bump on my forehead', 'The wind blows my hair about'. Examples of the second kind are: 'I see so-and-so', 'I try to lift my arm', 'I think it will rain:, 'I have a toothache'. (Wittgenstein, 1958, 66–67)

As the context makes clear, Wittgenstein is distinguishing two types of statement on the basis of the grounds on which they are made. The second group (the uses of "I" as subject) are all based on introspective awareness. Wittgenstein goes on to reframe the contrast in terms of susceptibility/immunity to a certain type of error —error of misidentification. The passage continues:

> One can point to the difference between these two categories by saying: The cases of the first category involve the recognition of a particular person, and there is in these cases the possibility of an error, or as I should rather put it: The possibility of an error has been provided for [...] It is possible that, say in an accident, I should feel a pain in my arm, see a broken arm at my side, and think it is mine when really it is my neighbor's. And I could, looking into a mirror, mistake a bump on his forehead for one on mine. On the other hand, there is no question of recognizing a person when I say I have toothache. To ask 'are you sure that it's you who have pains?' would be nonsensical. (Wittgenstein, 1958, 67)

Although it was plainly not Wittgenstein's intent to do, it will be helpful for the following to move from the formal mode to the material and talk, not about statements in which "I" is used as subject/object, but rather about the self-as-subject and the self-as-object. The idea is that judgments in which "I" is used as object, are judgments made on the basis of experiences in which the (bodily) self is presented as an object (and in which another object could mistakenly be presented). This bodily self is the self-as-object. Judgments in which "I" is used as subject are judgments made on the basis of forms of awareness in which the self-as-object does not appear in ways that can potentially give rise to errors of misidentification.

It is not hard to make sense of what the self-as-object might be, on this way of looking at things. But making sense of the self-as-subject is trickier. Wittgenstein himself warns us against one way of proceeding, which would be to "reify" the self-as-subject:

> We feel then that in the cases in which "I" is used as subject, we don't use it because we recognize a particular person by his bodily characteristics; and this creates the illusion that we use the word to refer to something bodiless, which, however, has its seat in our body. In fact, *this* seems to be the real ego, the one of which it was said, "Cogito, ergo, sum". (Wittgenstein, 1958, 69).

There is probably no danger of us making *that* mistake. However, there is another, much more subtle mistake that might be made here (and which, in fact, Wittgenstein himself may have made). The mistake emerges if one explains the

self-as-subject through the concepts we have been discussing of elusiveness and the zero-point. On this way of considering things, the self-as-subject is, so to speak, the vanishing point of thought and experience – and hence not as something that is, or can be, experienced. Once that initial step has been made, then we find ourselves with a stark and unattractive disjunction. Our experience of the self is either purely objectual, on a par with our experience of any other object in the world, or it is, strictly speaking non-existent.

This stark and unattractive disjunction is most clearly articulated by Wittgenstein himself in the 5.6 sections of the *Tractatus*, particularly in 5.631 and 5.632:

> 5.631
>
> There is no such thing as the subject that thinks or entertains ideas.
>
> If I wrote a book called *The World as I found it*, I should have to include a report on my body, and should have to say which parts were subordinate to my will, and which were not, etc., this being a method of isolating the subject, or rather of showing that in an important sense there is no subject; for it alone could *not* be mentioned in that book. –
>
> 5.632
>
> The subject does not belong to the world: rather, it is a limit of the world.

For the early Wittgenstein, therefore, there is either the self-as-object or. . . a zero-point. What motivates this picture is the assumption that the "I"-as-subject cannot appear in the content of perception. Once that assumption is made, it is an inevitable consequence that the self-as-subject is taken to be elusive (a zero-point).

But the assumption is mistaken. That it is mistaken is a consequence of various features of bodily awareness and exteroceptive perception that I have been emphasizing for a long time, certainly since *The Paradox of Self-Consciousness* in 1998, although I have only just realized the connection to Wittgenstein's distinction. The key point to recognize is that, *pace* Wittgenstein and those who think like him, the line between the self-as-subject and the self-as-object does *not* map onto the line between what can be experienced and what cannot. Rather, the distinction between self-as-subject and self-as-object can be drawn *within the content of perception*.

Both exteroceptive perception and bodily awareness contain multiple forms of self-specifying information that present the body in a way that clearly distinguishes it from all other physical objects. Since I have discussed them extensively elsewhere, here I will simply give a list, with the briefest of commentaries. They fall into three groups. The first comprises different types of information about one's own spatial orientation and trajectory that are given by high-level properties of outward-facing perceptual experience. These have been most extensively studied within the ecological approach to perception pioneered by the psychologist J. J. Gibson, who drew attention to multiple forms of self-specifying information available in the optic flow.[14] The second group is composed of the multiple forms of self-specifying information implicated in the ability to navigate the environment.[15] The third group contains all the different ways we have of finding out about the configuration, homeostatic status, and orientation/trajectory of our bodies through proprioception, kinesthesia, somatosensation, and so forth. These forms of information are distinctive, not only in that they uniquely provide information about one's own body, but also in that they underwrite a distinctive type of spatial phenomenology (referenced by both Gallagher and Serrahima). We experience the space of the body in a fundamentally different way from how we experience non-bodily space.[16]

---

[14] See Gibson (1966, 1979), with the implications for philosophical discussions of self-consciousness explored in Bermúdez (1995, 2002, 2003a), reprinted as Essays 2-4 in Bermúdez (2018e).

[15] On which see Bermúdez (1995) and, in connection with first-person thought and language, Bermúdez (2017a).

[16] See Essays 5-9 in Bermúdez (2018e), including versions of papers originally published as Bermúdez (2005c, 2011a, 2015, 2017a, 2018c).

Going back to the key passage from the Blue Book, Wittgenstein's criterion for distinguishing statements involving "I"-as-subject appears to be that they are immune to error through misidentiIication (relative to the first-person pronoun) as a function of the grounds on which they are made. Since, as I and others have argued at length, statements/ judgments made on the basis of the various types of self-specifying information listed in the previous paragraph are typically immune to error in the required sense, it follows that these are statements/judgments involving "I"-as-subject, in Wittgenstein's sense, and hence (in the material mode) that they are statements/judgments about the "I"-as-subject. In this sense, therefore, the self-as-subject is no less part of the world than the self-as-object.

Does this mean, though, that we should reject the basic idea of elusiveness, the idea that the world is experienced from an elusive zero-point that is not itself experienced? Not at all. But we do need to rethink how the elusiveness thesis is formulated. It seems to me that there is a single basic fact underlying both the thesis of the zero-point and the elusiveness thesis. This is the fact that no act of cognition or act of perception can be fully reflexive.[17] There is always an epistemic/perceptual distance between the *act* of cognition/perception and the *object* of cognition/perception.

If we think about elusiveness in this way, it comes out as a corollary of what might be termed the *non-reflexivity of consciousness*. This means, among other things, that we need to be careful about how the elusiveness thesis is formulated. As standardly formulated, the elusiveness thesis is a thesis about the self. This certainly seems to have been Hume's thought, judging by the following much-quoted passage:

> For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I can never catch myself at any time without a perception, and can never observe anything but the perception.[18]

It seems to me, though, that the phenomenon to which Hume is pointing is really a special case of a more general phenomenon. What carries the explanatory weight here is the assertion that: "[I] can never observe anything but the perception." That in turn seems to make most sense as a claim about irreflexivity: I can never observe anything but what is perceived (or thought, or experienced, or. . .). *A fortiori*, I cannot observe either the act of perception (or thought, or experience, or. . .) or the subject of perception (or thought, or experience, or. . .). This is almost certainly *not* what Hume himself was trying to maintain in this part of the *Essay*. He is generally thought, after all, to be arguing for a form of either eliminativism or reductionism about the self. But for those not wanting to accompany him down that road, it seems a more palatable option.

On this way of thinking about elusiveness, it is misleading to talk about the elusiveness of the *self*. It is true that, in any given case of introspection, the subject of what is being introspected cannot itself be introspected. But that is not a deep fact about the self. It is simply a byproduct of the basic fact that mental acts are not, and cannot be, reflexive. In this sense, therefore, elusiveness is not at all in tension with the idea that the self is directly perceived in bodily awareness (and indeed in exteroceptive awareness). In fact, elusiveness applies no less to experience of the self-as-subject than it does to experience of anything else, including the self-as-object and ordinary, common-or-garden physical objects.

Interpreting elusiveness in this way helps us to see that elusiveness holds for ϕ-ownership as well as for ψ-ownership – and for the same reason, which has everything to do with the irreflexivity of consciousness and nothing to do with the metaphysics of the self or the epistemology of self-awareness. Moreover, it illustrates that (although he might well not have accepted my reasoning) Husserl was right to characterize the phenomenology of bodily experience as being *from* a zero-point. And finally, it shows that, contrary to Serrahima's thought-provoking analysis, we can, and indeed should, give an account of the phenomenology of embodiment that incorporates both the idea that the self-as-subject is directly presented in experience and the idea that the body is experienced from an elusive zero-point.

---

17 At least, I take it to be a fact. For a different view, see Caston (2002).

18 Hume, *Treatise of Human Nature* (Selby-Bigge, 1888, p. 252).

Still, the conclusions of the previous section notwithstanding, important questions remain about the relation between elusiveness and phenomenology for anyone who thinks that we need an integrated account of ψ-ownership and ϕ-ownership. Whereas in the case of ϕ-ownership, it is elusiveness that seems to pose difficulties, in the case of ψ-ownership it is phenomenology. In particular, one might ask:

1) Is there a role for phenomenology in an account of ψ-ownership ?
2) Suppose that bodily self-ascriptions do indeed involve taking bodily experiences at face value, does anything like this hold in the case of introspection?

I will end this paper with some brief and tentative thoughts about the possibility of answering both these questions in the affirmative.

Unsurprisingly, much depends on how one understands the basis for psychological self-ascriptions, and how one thinks about introspection. In the following, I will consider just the narrow case of beliefs and the family of transparency accounts of self-knowledge.

Transparency accounts of self-knowledge are often traced back to some suggestive remarks in Gareth Evans's *The Varieties of Reference*, where Evans observes, surely correctly, that we typically answer the question of whether we believe that *p*, not by some sort of internal self-scrutiny, but instead by considering the question of whether *p* is actually true.[19] Generalizing this initial thought, Alex Byrne has identified what he terms the doxastic schema as underwriting transparency accounts of self-knowledge (Byrne, 2018). The doxastic schema identifies inferences of the following form as providing warrant for self-ascriptions of belief.

P

I believe that *p*

The doxastic schema seems to involve a "taking at face value" that is directly analogous to that which occurs in bodily self-ascriptions. To apply the doxastic schema is to arrive at self-ascriptions of belief by taking one's beliefs at face value. There would seem to be a direct parallel between the way in which one arrives at the bodily self-ascription "I am in front of the cathedral" taking one's visual experience at face value and the way in which one arrives at the psychological self-ascription "I believe that the cathedral was built in the fifteenth century" by taking at face value one's belief that the cathedral was built in the fifteenth century. There are prospects then for answering the second question in the affirmative.

A final, even more tentative suggestion. There are ways of developing the transparency view that point towards an affirmative answer to the first question. On the bypass model proposed in Fernandez (2013) (and which seems very much in the spirit of Evans's original remarks), the grounds for one's self-ascription of the belief that *p* are exactly the grounds that one has for believing that *p*. If that is right then it would seem to give one way in which psychological self-ascriptions (and hence ψ-ownership, more generally) can have a phenomenological dimension (at least in the case of beliefs). In many cases, for example, the grounds for psychological self-ascriptions will include psychological states that directly present the self-as-subject. This is most obviously the case for beliefs about one's own psychological states that involve simply taking those states at face value. If I come to the conclusion that I believe that I am in front of the cathedral by applying the doxastic schema together with my visual experience of the cathedral in front of me, then (on the bypass view) my psychological self-ascription is based on a presentation of the self-as-subject. But it will hold also for any psychological self-ascription for which the evidential basis involves a comparable presentation of the self-as-subject – suppose, for example, that I attribute to myself the belief that I am in Seville because, among other things, I see Seville Cathedral in front of me.

[19] See, for example, Evans (1982, p. 225).

Plainly there is much more to be said here, but hopefully this discussion of these two stimulating papers by Gallagher and Serrahima has pointed towards the possibility of developing an integrated account of ψ-ownership and ϕ-ownership that incorporates both the idea that we experience the world from an elusive zero-point and the idea that we are directly presented with the self-as-subject in bodily awareness and exteroceptive perception.

4. Inner speech and intentional ascent: Reply to Bernués and Jorba

In 'Bermúdez's view on inner speech: A critical assessment', their contribution to this volume (Bernués and Jorba 2026), Daphne Bernués and Marta Jorba develop a number of very interesting arguments, and I thank them for engaging so carefully with the arguments I developed in *Thinking without Words* (Bermúdez, 2003c) and then subsequently in 'Inner speech, determinacy, and thinking consciously about thoughts' (Bermúdez, 2018d). In this brief response I will leave the first parts of their paper relatively untouched, making only a few remarks that will hopefully clarify my overall project and perhaps go some way towards blunting their objections. My main focus will be on the dilemma that they raise for me in the fourth part of their paper. I present the dilemma in §4.2 and then discuss each horn in turn in §§4.3 and 4.4. Some of what they say there is well taken, and I am grateful to have the opportunity to revise, and hopefully strengthen, my position in a way that allows me to steer through the dilemma.

4.1. What do the data show?

One of the central claims of *Thinking without Words* is that, although the cognitive abilities of prelinguistic infants and nonlinguistic animals are surprisingly rich, there are principled limits to how far they can extend. In particular, so I claimed, certain types of thinking require linguistic vehicles. These types of thinking all involve what I call intentional ascent – that is, they are types of thinking that take thoughts as their objects. Examples that I discussed there include propositional attitude mindreading, higher-order desires, thoughts that involve the application of logical operators and quantifiers, and the types of reflexive thinking that involve monitoring and evaluating the evidential and justificatory relations between thoughts (or between thoughts and other mental states, such as perceptions). All of these forms of intentional ascent require and presuppose semantic ascent —we can only think about thoughts as linguistically expressed ("clothed in the garb of language", as Frege famously put it). In 'Inner speech, determinacy, and thinking consciously about thoughts' (Bermúdez, 2018d), I developed this line of argument further in the context of inner speech. There are many forms of semantic ascent. I can reflect on thoughts as written down or as spoken out loud —by myself or by someone else. But inner speech is surely a key tool for semantic ascent, or so I argued there.

The third section of Bernués and Jorba's paper present a range of empirical results that, they claim, show my position to be "empirically untenable". The results that they cite are certainly interesting and (largely) significant. This is not the place to go through them in detail and so I confine myself to some more general remarks.

First, some methodological comments. I have little sympathy for *a priori* "refutations" of empirical data and have always tried to develop my views in a way that is sensitive to important results in the social and behavioral sciences. But, at the same time, and to use Quine's famous metaphor of science as a force-field, philosophy is at some distance from the empirical boundary of the field of science. One cannot simply read philosophical results from the data. Instead, and as I argued in 'The force-field puzzle and mindreading in non-human primates' (Bermúdez, 2011c), we need a more nuanced view of the relation between data and observation, on the one hand, and theorizing on the other. In particular, we need to develop (simultaneously and in conjunction with each other) conceptual arguments and interpretative frameworks for making sense of experimental evidence.

The point of mentioning this is that, for each of the types of intentional ascent identified earlier, I have argued that there exist more primitive analogs that *are* available at the non-linguistic level. The focus of the force-field paper is

on the distinction between propositional attitude mindreading, which does involve intentional ascent, and perceptual mindreading, which does not. My conjecture there was that the experimental evidence typically cited to support claims about mindreading in nonlinguistic creatures is best understood as illustrating perceptual mindreading, rather than propositional attitude mindreading. I would make similar conjectures for apparent counter-examples to my claims about the language-dependence of other forms of intentional ascent. So, for example, in *Thinking without Words* I argued that nonlinguistic creatures can reason in ways that are structurally similar (in limited domains) to classical inference patterns, such as disjunctive syllogism, even though they cannot (if my arguments are sound) reason with logical operators. What I called proto-logic extends to negation and to universalizing thoughts —just not to negation as an operator on propositions and to universal quantification understood *in sensu compositu*.[20]

This should give pause for thought in interpreting the experimental data. It is far from obvious to me that the experiments that Bernués and Jorba cite as examples of how one can have different types of intentional ascent without inner speech or public language really are illustrations of intentional ascent, properly speaking, rather than the more primitive analogues that I claim not to involve intentional ascent. At any rate, more work needs to be done here.

4.2. The dilemma

In their section 4, Bernués and Jorba raise an interesting challenge to my position, in the form of a dilemma. They state the dilemma as follows:

> To sum up, we are in front of a dilemma. On the one hand, if thought content is determined subconsciously, then the role of inner speech becomes superfluous in the revelation of canonical structure (condensed and expanded forms would do equally well, and presumably forms of unsymbolized thought too). On the other hand, if expansion is decisive for revealing the canonical structure of thought, formalization (or any other sophisticated form of inner speech) would be a secure case of doing the job, but at the cost of distancing too much from how typically inner speech is experienced. (Bernués and Jorba 2026, p. 225)

The starting point for the dilemma is what they (and many others) take as a basic datum about inner speech, namely, that it is condensed and fragmented. The point is nicely made in a 2010 paper by Martínez-Manrique and Vicente:

> The Vygotskyan approach to inner speech regards it as a phenomenologically degraded form of our own talking, which is also syntactically simplified. In this view, inner speech not only lacks pitch and volume but also appears typically in subsentential linguistic items. So inner speech is not experienced in the way of the internal monologues that classical novels depict but in the fragmentary way exemplified by the opening of Samuel Beckett's "The Unnamable": "Where now? Who now? When now? Unquestioning". Our "outer talk" also often has subsentential utterances but our inner talk often seems to be full of expressions like 'ah', 'yes', 'not this way', 'where the hell?' and 'the meeting!' [...] The meaning of those linguistic items is typically clear to us but they might be to a large extent unintelligible to others if we uttered. (Martínez-Manrique & Vicente 2010, p. 142)

I am not convinced that these Beckett-like subsentential utterances, although they plainly exist, and perhaps are even widespread in inner speech, have much of a role to play when inner speech is being used for intentional ascent (and, in line with the points made in the previous section, my view is that the types of thinking that involve inten-

[20] The distinction alluded to here is between general beliefs *in sensu diviso* (e.g., when I believe, of every tiger, that it is dangerous) and general beliefs *in sensu compositu* (e.g., when I believe that every tiger is dangerous). Intentional ascent is involved only in those general beliefs where the universal quantifier is in the scope of the belief operator (i.e. *in sensu compositu*).

tional ascent are themselves far less widespread than they are commonly taken to be[21]). But still, Martínez-Manrique and Vicente make a good point and, as they go on to argue, it presents a *prima facie* difficulty for another aspect of my view:

> Let us remark again that the linguistic items that appear to conscious thought can be often condensed and subsentential. These simple linguistic items have to be "interpreted as expressing a proposition" much in the way Bermúdez suggests that maps have to be. Thus, there is, prima facie at least, no difference between an analogue vehicle and NL in this respect. If NL [natural language] can help us have a second-order dynamics, despite its inexplicitness, there is no reason why analogue formats cannot. (*Ibid*, p. 146)

Setting aside the issue about maps, this is what Bernués and Jorba home in on to set up their dilemma.

The dilemma emerges when we ask: How do we get from condensed utterances in inner speech to the type of articulation of a thought that is required for intentional ascent? My argument for the language-dependence of intentional ascent (well summarized by Bernués and Jorba at pp. 222-223 of their paper) rests upon the need for thoughts to be represented in a way that reveals their structure. But, if Martínez-Manrique and Vicente are right, then condensed utterances in inner speech would need to be interpreted (or at any rate, transformed in some manner) in order for this to be possible. The dilemma emerges when we ask whether this interpretation/transformation takes place unconsciously at the subpersonal level, or consciously at the personal level. For the first horn, assume that it takes place subpersonally. Then, according to Bernués and Jorba, inner speech would be superfluous. The representation of a thought required for intentional ascent would be whatever the output is of that subpersonal interpretation/ transformation, and it would certainly not be conscious and available to reflective evaluation and the other operations of intentional ascent. Suppose then (for the second horn) that it is conscious. Bernués and Jorba claim that in this case inner speech would also be superfluous, but for a different reason. The most perspicuous representation of the structure of an utterance in inner speech would most likely not be in the form of anything that looks like a natural language – as opposed to being in some artificial language designed for perspicuously representing the structure of natural language (first-order logic, for example).

4.3. The first horn

To start with the first horn of the dilemma, Bernués and Jorba elide a distinction that I make between two different ways of developing what I call the *determinate translation hypothesis* —which is the hypothesis that the objects of intentional ascent must be semantically determinate in a way that involves "translating" semantically indeterminate inner speech utterances into mental representations that resolve ambiguity and context-sensitivity into contents sufficiently determinate for intentional ascent.[22] It is common to both strategies that indeterminacy is resolved at the subpersonal level, through mechanisms and processes that are in principle inaccessible to conscious awareness. The difference comes when we ask when and how the resolution of indeterminacy takes place. On the first view, it takes place *downstream* of the inner speech episode. That is to say, a semantically indeterminate inner speech utterance is the input to some sequence of operations that yields a semantically determinate (or, at any rate, significantly less indeterminate) mental representation that can then be the object of (input to) further operations corresponding to what I am terming intentional

---

[21] I have argued in several places, for example, against the widely held view that all but the simplest forms of social interaction and coordination depend upon propositional attitude mindreading. Propositional attitude mindreading (of the type that involves intentional ascent) is relatively infrequent, applied when simpler mechanisms for social understanding and social coordination reach their limits. See, for example, Bermúdez (2003b, 2018b).

[22] For the record, I am fairly sympathetic to neo-Wittgensteinian arguments that the determinate translation hypothesis is completely misconceived. Indeterminacy is resolved by context and social practices, not by psychology. As memorably put in the *Philosophical Investigations*, "If God had looked into our minds he would not have been able to see there whom we were speaking of" (Wittgenstein, 1953, p. 217). But for present purposes it makes sense to meet Bernués and Jorba on their own ground.

ascent. Bernués and Jorba are surely correct that this way of resolving indeterminacy would make inner speech episodes redundant, since the processes of intentional ascent completely bypass the inner speech episode.

There are independent reasons, though, for being skeptical of this approach, which seems to run into a dilemma of its own. Suppose we ask whether the mechanisms of intentional ascent that operate on these mental representations with (relatively) determinate contents do so consciously or not. If they are not conscious, then the phenomena under consideration are not the phenomena that I was trying to explain. The basic starting point for my argument is that the paradigm cases of intentional ascent are conscious mental acts in which a target thought is taken as the object of a second–order thought (this is premise 1 in the argument given on p. 222 of Bernués and Jorba's paper). There might or might not be such a thing as non-conscious intentional ascent, but even if there is, it is definitely not what I am talking about. So, let us consider then the second horn. If the operations of intentional ascent are consciously accessible, then so too must be the inputs into those processes, namely, the disambiguated mental representations. But now there seem to be two consciously accessible items in a given episode of intentional ascent —the original inner speech utterance, on the one hand, and the mental representation that gives its (relatively) determinate content, on the other. This seems a somewhat extravagant account of what is going on when, for example, I run through an argument in my head to check it for validity.

Is it really plausible that, in addition to being conscious of the argument as represented by a sequence of sentences in inner speech, I am also conscious of another sequence of representations that represents the same argument, just in a less indeterminate way? Perhaps the idea is that the inner speech episode should be understood purely acoustically, so that the only content-bearing representations are the outputs of the process of determinate translation? But that seems, if anything, even less plausible. Only very rarely, if ever, do we have inner speech experiences that are, so to speak, uninterpreted. In this respect, the phenomenology of inner speech is very similar to that of outer speech – what we hear comes to us as always already interpreted. There are, of course, occasions in ordinary conversation where we do not understand what is being said, and we have to guess. In *those* case there is a sense in which we can plausibly be said to hear an uninterpreted utterance in conjunction with an interpretation. But not only are they outliers (and really we have something more like a conjecture than an "interpretation"), but it is most unclear that anything analogous ever occurs in inner speech.

Fortunately, there is another way of thinking about how indeterminacy might be resolved at the subpersonal level that is better aligned with the phenomenology of inner (and outer) speech. This would be to think of the resolution of indeterminacy as taking place *upstream* of the inner speech episode. On this model, we are not aware, in an inner speech episode, of something that is systematically ambiguous and indeterminate because what is accessible to consciousness reflects and incorporates the results of subpersonal information-processing. Because there is no need for further processing, this account will obviously avoid the puzzles just discussed and does justice to the widely-noted observation that we are rarely in doubt as to the content of inner speech episodes. This aspect of inner speech is highlighted in the model proposed in Carruthers (2018), which can be seen as an example of this second type of approach.[23] The approach helps also with the fragmentary nature of many inner speech episodes, as highlighted by Martínez-Manrique and Vicente. The following passage from Carruthers brings both of these points together:

> We are often at a loss to know what others have just said. But it seems we rarely, if ever, have similar doubts about our passing thoughts. This is true despite the fact that inner speech can often seem highly fragmentary, consisting of just a word or phrase. And in many of these cases, it seems that the propositional content we hear ourselves express isn't obviously given by the surrounding semantic context. (Contrast a case where one hears someone respond to the question "Will it

[23] See particularly section 1.2.4. Carruthers does think that there are separate mechanisms for the production and comprehension of inner speech episodes, but from a phenomenological perspective, if I understand his position correctly, what we experience is the result of the comprehension mechanism, not input into it.

> rain today?" with a simple "Yes". Here there is no difficulty explaining how one hears the latter as expressing the content, "It will rain today." No doubt similar phenomena can occur in inner speech also. But many cases of fragmentary inner speech with determinate heard content don't seem to fit this sort of profile). (Carruthers 2018, p. 45)

This second *upstream* model of how semantic indeterminacy is resolved has, therefore, a number of theoretical advantages. Nonetheless, Bernués and Jorba may feel that these theoretical advantages do not include answering the key claim of the first horn of their dilemma! After all, their objection is that allowing for the subpersonal resolution of semantic indeterminacy makes inner speech *redundant*, and this objection would seem to hold whether the subpersonal resolution takes place upstream or downstream of the inner speech episode. Either way, the inner speech episode is not involved.

In response, I should make clear that I do not believe I ever said that inner speech was involved in semantic disambiguation. My claim is about intentional ascent and the role that inner speech can play in making first-order thoughts available for second-order reflection. Even though I have granted (for the sake of argument) that intentional ascent operates on semantically determinate thoughts, that concession by no means commits me to holding that semantic indeterminacy is resolved by and in inner speech. Inner speech is a powerful tool, but I do not think that it is *that* powerful. The operation of resolving semantic indeterminacy seems to me to be a completely different operation from the operation of intentional ascent. The second might entail the first, but I do not think that the first entails the second.[24]

4.4. The second horn

This discussion brings us to the second horn of Bernués and Jorba's dilemma. Even if we agree with Carruthers that we are highly reliable in attaching semantic contents to inner speech episodes (or, as I would prefer to put it, that we experience inner speech episodes with determinate contents), it might still be objected that these are not the right sort of contents to be the object of reflective thinking about thinking. Recall that a central premise in my argument is that a thought can only be the object of a second, higher-order thought if it is represented in a way that reveals its canonical structure. But can we be so confident that, even if semantic indeterminacy has been resolved, a natural language utterance in an inner speech episode represents the content of a thought in a way that reveals its canonical structure?

Frege himself thought that natural language was an imperfect guide to logical structure, which he took to be better revealed in a logically perspicuous formal language. Frege's principal concern was with mathematical reasoning, not natural language, but the Davidsonian program in the philosophy of language was intended to carry forward the broadly Fregean project of attempting to regiment natural language in a semantic theory grounded in first-order logic. Parallel projects in linguistics and formal semantics have turned to different types of formal theory to represent natural language structure (and thereby the structure of the thoughts that those natural language sentences express). The details are not important for present purposes. What matters is the common thread to all these research projects, which is that they all take canonical structure to be represented in formats other than natural language sentences. So, if something like that common thread is true, Bernués and Jorba argue, then natural language seems to have no work to do in enabling thinking about thinking. The objects of intentional ascent are not natural language sentences, but rather sentences in first-order logic, Montague semantics, or some other formal representation.

---

[24] There is an important issue here that needs further discussion. Obviously, if resolving semantic indeterminacy *required* some sort of process of hypothesis formation and testing than it *would* involve a form of intentional ascent. Perhaps something like this thesis is in the background of Bernués and Jorba's discussion? It certainly requires further discussion. For now I will confine myself to observing that, in the case of spoken language, we often disambiguate semantic indeterminacy by thinking about the world, not by thinking about words. How do I disambiguate "Mary had a little lamb", to take a well-known example? Typically, by attending to contextual clues in the conversation and/or the environment, not by forming semantic hypotheses. Or at least not by *consciously* forming semantic hypotheses, and, as mentioned above, even if there is such a thing as subpersonal intentional ascent, that is not what I am talking about.

There are (at least) three reasons for not being satisfied with this line of argument, however. The first is that it over-interprets the notion of structure required for my argument. The central claim in my argument is that natural language sentences reveal canonical structure in a way that imagistic representations do not. The expression "canonical structure" may be muddying the waters here. Let me make the point without it. Suppose that I am attributing a thought to someone – the thought, say, that the church is to the left of the Chinese restaurant. This is an exercise in propositional mindreading, not perceptual mindreading. I am not visualizing how the world might look to someone standing in front of the Chinese restaurant and seeing a church to the left. I am characterizing them as thinking about the arrangement of buildings. So, for reasons familiar from discussions of compositionality, the Generality Constraint, and so on, I am characterizing them as exercising a complex of abilities that the thinker could exercise in different contexts (and indeed, were they not able to exercise those abilities in different contexts then it would not be appropriate to attribute a thought to them in the first place). To be able to think that the church is to the left of the Chinese restaurant, someone needs to be able to think other things about the church and the Chinese restaurant, to think for another pair of objects A and B that A is to the left of B, and so on. This is the sense in which the object of intentional ascent needs to be structured for the purpose of attributing propositional attitudes – it needs to be represented in a way that distinguishes between these different abilities. A natural language sentence actually performs this function rather well —no worse, one might think, than a representation in first-order logic or NL (and probably better, I am inclined to say!).

The second problem is that the objection significantly underestimates what can be done within natural language. To see this, consider a different type of intentional ascent, namely, that involved in logical thinking – i.e., tracing out the inferential connections between a set of premises and a conclusion. It is true that natural language can be deceptive, in well-known ways. Not all names refer, for example. Definite descriptions behave differently from proper names. Natural language is highly prone to quantifier shift fallacies and other forms of logical ambiguity. And it is also true that formal languages can disambiguate in helpful ways. Changing the order of the quantifiers in a formal representation of "every boy loves some girl" makes the scope ambiguity very clear. Likewise, the existential assumption in a definite description (if indeed there is one) can be brought out clearly in a first-order representation, as it is in Russell's theory of descriptions. But still, formal representation is hardly *necessary* in either of these cases. Medieval logicians had no problem explaining (in Latin) the difference between there being a single donkey that every farmer beats and every farmer beating some donkey, but not necessarily the same donkey. And Russell gives us a pretty accurate account in English of how he thinks we should understand expressions such as "the present king of France". Finally, to take another language, Frege offers penetrating accounts in his *Logische Unterschungen* of how natural language can deceive us with respect, for example, to the logical properties of negation. But his analyses are offered in German, not the formal language of the *Begriffsschrift*, which he used primarily for representing mathematics.

A final problem with this line of argument is that it is a *non sequitur*. Even if one were unconvinced by the previous two points and maintained that ultimately intentional ascent really does require a formal representation of structure, the fact remains that one has to start from somewhere to get to that formal representation. And what could that starting-point be but a natural language sentence? Formal semantics is, after all, formal semantics – that is to say, a semantic theory for natural languages. Even if one thinks, as many do, that the structure of a thought is most perspicuously represented in some formal language, that thought needs first to be expressed in a natural language sentence before the formal analysis can get started. I should emphasize again that the intentional ascent I claim to be language-dependent is that involved in conscious forms of higher-order thinking, and I am not sure that I understand what it would mean to say that the objects of conscious thinking about thoughts are formulae in a logical language, unless those formulae are being thought about as representations of the structure of a natural language sentence.

* * *

These brief remarks will hardly settle the issue, but they will hopefully make clear how I would respond to the dilemma that Bernués and Jorba raise for my position. I end by thanking both authors for developing a very interesting line of argument from which I have learned considerably.

5. *Preferences, frames, and reasons: Reply to Dohrn and Fisher*

In *Frame It Again: New Tools for Rational Thought* (Bermúdez, 2020) and subsequently in a 2022 target article in *Behavioral and Brain Sciences* (Bermúdez, 2022a, 2022b), I explored the role of frames and framing in practical reasoning. I argued, in opposition to decision-theoretic orthodoxy, that framing phenomena and framing effects are not experimentally induced deviations from ideal rationality, which is how they are standardly treated in a long psychological tradition stemming from the landmark studies reported in, e.g., Tversky and Kahneman (1981, 1986) and Kahneman and Tversky (2000). Instead, we should be thinking about framing as an integral and inescapable part of practical reasoning.

It is, I maintain, a mistake to think of preferences as invariably extensional, so that one must prefer one object or state of affairs to another *simpliciter*, irrespective of how they are described, apprehended, or given. This is in part a descriptive mistake. Preferences just do not work like that. Preferences are frame-relative. But it is also a normative mistake, in two different ways. First, if we define a framing effect to occur when a single thing is knowingly valued differently depending on how it is framed, then I argued that framing effects are not necessary irrational. There are rational framing effects. Moreover, and this is the second point, rational framing effects have the potential to play very important roles in a wide range of reasoning situations, including ones that seem *prima facie* to be completely intractable.

Daniel Dohrn (2026) and Sarah Fisher (2026), in their stimulating and constructive contributions to this special issue, press me on both the descriptive and normative dimensions of my views on framing. Dohrn is skeptical of my thesis that preferences are frame-relative, and he is also doubtful about the possibility of attributing what I term quasi-cyclical preferences (which is the preference structure that I identify in rational framing effects). He takes issue, in short, with how I describe the landscape of choice. He is also doubtful that any framing effects, however they are described, could be rational, and he is unconvinced by my arguments that rational framing effects can be deployed to solve problems in game theory and self-control. Fisher is equally unconvinced by the normative dimension of my views on framing, but she focuses her attention on the final two chapters of *Frame It Again*, where I develop a model of frame-sensitive reasoning that is intended to provide a way to overcome what I call discursive deadlock (which arises, often, on so-called "values issues", when standard tools and techniques for dispute resolution and collective decision-making fail). Fisher thinks that I am overly optimistic about the power of frames, because I over-estimate the role of frames in providing reasons for action (and more generally in reasoned debate), while underestimating the role of frames in signaling identity and affiliation.

In the following I begin in §5.1 by outlining how I see preferences and why I think that preferences can be frame-relative. Against that background, §5.2 then engages with Dohrn's discussion of my claim that preferences are frame-relative in a way that opens the door to what I term quasi-cyclicality, taking this as an opportunity to explore the nature of preference more deeply. In §5.3 I turn to Dohrn's objections to the thesis that quasi-cyclical preferences can be rational. Finally, §5.4, explains why I am more sanguine than Fisher about the prospects for frame-sensitive reasoning, although I certainly take on board her insightful discussion of how frames can (in the contexts she discusses) serve a signaling function that actually reinforces discursive deadlock, rather than dispelling it.

5.1. Frame-relative preferences

The foundational principle of decision theory is the *expected utility principle*, which is an equation with preferences on one side and probabilities and utilities on the other. The principle says that an agent prefers *a* to *b* if and only if the ex-

pected utility of *a* is greater than the expected utility of *b*, where the expected utility of an action is given by the sum of the utilities of the different possible outcomes of the action, each weighted by the probability that it occurs. The mathematical basis for the expected utility principle is typically what is called a representation theorem, proving that, provided an agent's preferences are well-behaved and consistent in the sense of conforming to a small number of basic axioms, utility and probability functions can be defined in such a way that the agent comes out as an expected utility maximizer. As far as the representation theorem is concerned, preferences are basic, with probabilities and utilities constructed from them. Stepping back from the representation theorem, the doctrine of revealed preference takes preferences to be directly revealed in choice, so that choice and preference are equivalent. On this way of considering the basic theoretical notions of decision theory, the direction of explanation goes from agents' preferences (as operationalized in their choice behavior) to their probability and utility functions.

As many have noticed, it is very tempting to see decision-theoretic probability and utility functions as regimentations of the folk psychological notions of belief and desire. Interpreted thus, agents' probability functions represent their degrees of belief in different possible ways the world might be, while their utility functions represent how they value the possible outcomes in those different possible ways the world might be. On this view, the expected utility principle is a formal analog of the intuitive principle that people typically do what they believe will best satisfy their desires, and decision theory comes out as a formal treatment of belief-desire psychology.

Interpreting decision theory in this manner, however, requires us to treat preferences very differently. We certainly cannot take preferences to be revealed in choice. Agents typically choose one course of action over another, and in most cases the action they choose will be the one that they prefer. But of course any satisfactory model of psychological explanation will have to allow for a gap between preference and choice, in order to accommodate lack of self-control, weakness of will, and similar phenomena. Moreover, we cannot take the structure of an agent's preferences as given (or as simply revealed in choice). If we are in the business of psychological explanation then we must treat preferences not just as psychological entities, but also as psychologically grounded. In other words, we have to explain why an agent has the preferences that they do.

The question of the psychological basis for preference has not been extensively explored within decision theory (or, for that matter, within philosophy), but to the extent that there is a standard view, it is that the process of explaining where preferences come from is essentially a representation theorem in reverse. That is to say, agents' preferences are to be derived from their beliefs and desires, on the basic assumption that agents prefer *a* to *b* when and only when they believe that *a* will satisfy their desires better than *b*. The basic starting-point for my discussion of frames and their role in practical reasoning is that this way of thinking about preferences is hopelessly impoverished. If we want to do justice to the complexities of practical reasoning then we need a more sophisticated model of how preferences are formed. It is through this more sophisticated model that frames enter the picture.

So – where do preferences come from? An obvious starting point is that preferences are grounded in valuations. We prefer what we value more. But where does value come from? Classical decision theory is, broadly speaking, Humean in spirit. That is to say, it takes practical reasoning to be purely instrumental. Agents' valuations (their desires, ends, utility assignments, pro-attitudes – whatever you want to call them) are taken as given, and the job of practical reasoning is to find means for satisfying those desires, or achieving those ends. Certainly, this way of considering valuations is very useful, and perhaps even necessary, for any systematic, formal investigation of the instrumental dimension of practical reasoning. But it is important to remember that it is a useful *fiction*. Desires, ends, utilities, and so on are not brute facts, like the color of one's eyes or the size of one's feet. They are the result of complex interactions between personal history, upbringing, social environment, and a wide range of psychological (and bodily) states.

We start to appreciate the role of frames in preferences when we appreciate, not just that valuations have this sort of complex aetiology, but also that they are neither constant nor homogenous. This variation arises through many different mechanisms, but for now we can focus simply on two. The first is how we perceive similarities. We often value things as a function of how similar they seem to other things that we value. But similarities are not simply given. The

similarities that we see between things are often determined by how we frame them. To take an example discussed in Bermúdez (2019b), genetically modified organisms (GMOs), can be viewed under a husbandry frame, or under a monopoly capitalism frame. Under the first frame, GMOs are similar to more "homely" types of selective breeding, and so one might value them (positively?) in line with one's valuations of traditional agricultural activities.

Under the second frame, in contrast, the similarities are to the ways in which multinational corporations can hold smaller businesses hostage, and are far more likely to be negative. Nor of course do these frames themselves come out of nowhere. To frame something a certain way is to focus on and foreground one of its aspects. The husbandry frame is a natural concomitant of foregrounding the complex processes by which GMOs are derived, while the monopoly capitalism frame emerges when one focuses on some of the commercial practices that surround GMOs (e.g., the fact that farmers are often contractually required to purchase new seeds each year).

Valuations are also driven by emotional engagement and, as with perceived similarities, how one engages emotionally with an outcome can vary according to how it is framed. To take an example discussed at length in Ch. 6 of *Frame It Again*, there are many different ways of framing climate change. One can think about climate change under a natural disaster frame, for example, highlighting impacts such as droughts, landslides, storms, and so on. This way of framing climate change can typically provoke strong emotional responses, as is well understood by many organizations engaged in climate-change-related fundraising. An alternative framing might be in terms of one's obligations to future generations. This framing is perhaps less emotive —or at least, associated with different and arguably less visceral emotional responses. Again, these two frames do not come out of nowhere. Each is the result of focusing and foregrounding a different aspect of what is indisputably a highly complex phenomenon. And framing the phenomenon one way rather than another will bring certain types of reason into the foreground, while making others recede into the background.

Someone who accepts the argument so far has already traveled most of the way towards frame-relative preferences and the possibility of quasi-cyclicality. The basic point is that preferences over outcomes can shift as a function of other psychological factors (of which we have considered two, but there are of course others). Since those preference-shifting psychological factors can themselves be shaped and transformed by how the relevant outcomes are framed, it follows that preferences can depend upon how outcomes are framed, and that is all that I mean by talking about frame-relative preferences.

In many cases, the frame-relativity of preferences is unlikely to make much, if any, practical or theoretical difference. This is so particularly in what Savage called *small* worlds (namely, decision problems that are simple enough to support standard idealizations about the foreseeability of all possible consequences, and so on). Decision theorists have tended to take the small world cases as paradigmatic, and so relatively little attention has been paid to the difficult cases where frame-relativity becomes salient. These difficult cases are ones where frame-relativity leads to what is sometimes called preference reversal, namely, when an agent's preferences over two possible actions changes as a function of changes in one or both actions. In certain circumstances, this can yield what I term *quasi-cyclical preferences*. Quasi-cyclical preferences arise when an agent *knowingly* frames a single option in two (or more) different ways. Relative to one framing he prefers that option to an alternative. But relative to a different framing, he prefers the alternative. Or so at least I claim. Dohrn, however, disagrees, as we will see in the next section.

5.2. Quasi-cyclical preferences?

Dohrn (2026) takes me to task, with some justification, for saying that there is no such thing as making decisions over a purely extensional opportunity set. On one way of reading that claim, it is stronger than I need and I welcome the opportunity to clarify the matter. Theorists of a broadly Fregean persuasion (among whom I count myself) typically think that propositional attitudes, including desire and preference, can be directed at objects only under certain modes of presentation (i.e., only as those objects are presented in a certain way). From that perspective, there cannot be a purely extensional opportunity set (or indeed, any purely extensional object of propositional attitudes), simply as a matter of

psychological fact. But of course, that is perfectly compatible with holding that, to all intents and purposes, we make decisions as if we were making them over a purely extensional opportunity set. What this means is that any apparent breach of extensionality is instantly corrected. Suppose, for example, that I prefer *a* to *b* and *b* to *c*, unaware that *a* and *c* are really the same option, presented in two different ways. Even the most extensionally-minded of decision theorists must accept that this can sometimes occur. But what they would insist is that anyone who discovers that *a* and *c* are really just different ways of presenting a single option will revise their preference structure and decide whether they prefer *a* = *c* to *b*, or vice versa. In other words, it is simply not possible to prefer *a* to *b* and *b* to *c*, if I know that *a* and *c* are really the same option, presented in two different ways.

In the first instance, this is a descriptive issue. The question is about the nature of preference, namely, whether preference is the sort of psychological state that would allow for the possibility of what I term *quasi-cyclical preferences*. My basic claim is that, thanks to the mechanisms sketched out above, quasi-cyclical preferences are perfectly possible. Two necessary conditions must be satisfied for quasi-cyclical preferences to occur. The first is that an option be framed in two (or more) different ways *at the same time*. Quasi-cyclical preferences are simultaneous. The second is that the agent must know that they are dealing with a single option framed in two different ways.

In *Frame It Again,* I discussed a range of decision problems in which I consider preferences to have a quasi-cyclical structure. These includes literary dilemmas (Agamemnon at Aulis, as portrayed by Aeschylus, and Shakespeare's Macbeth), illustrations from game theory (including the so-called Chicken game, as well as both the Prisoner's Dilemma and the Altruist's Dilemma), patterns of choice in problems of self-control, and a range of different so-called "values" issues. In his paper Dohrn offers an additional literary illustration, from Victor Hugo's *Les Misérables*. Although Dohrn himself does not think that this example should be described in the way that I want to describe it, it seems to me to fit my model rather well, and so I reproduce it with gratitude. (Dohrn marks frames with italics and initial caps, placing the actions that they frame in parentheses.)

> In volume one of Victor Hugo's *Les Misérables*, Jean Valjean is faced with a tragic choice: on the one hand, he may turn himself in to save a ruffian who faces an unjust punishment for minor offences committed by Jean Valjean himself in the distant past.
>
> On the other hand, he may keep up his disguise. Doing so would mean to continue serving a whole community that depends on his entrepreneurship and charity and in particular to save the innocent Fantine and her child from misery and death. There is no satisfying choice. Valjean turns himself in, and this is intimated to be the morally right choice, but the consequences are dire: the community is ruined and Fantine dies. Before turning himself in, Valjean spends the whole night describing the situation under different frames. This is a typical case of the sort Bermúdez has in mind. Under the frame *Serving the Community* (not turning oneself in) trumps *Saving the Ruffian* (turning oneself in). Under the frame *Taking Responsibility for Past Offences* (turning oneself in) trumps *An Innocent Being Punished instead of Oneself* (not turning oneself in). (Dohrn, 2026, p. 237)

Both conditions upon quasi-cyclical preferences seem satisfied. First, there are two options, such that Jean Valjean simultaneously prefers the first to the second under one frame and the second to the first under a different frame. Second, moreover, he is completely clear that these are two different ways of viewing a single option.

Dohrn disagrees, however. He raises two objections. The first has to do with the nature of agency, and specifically the relation between agency and acting upon a preference. He argues that, if things were as I describe them, then it would not be correct to describe Jean Valjean as *acting*. Dohrn's second objection focuses on how preferences are attributed, arguing that attributing to Jean Valjean a quasi-cyclical preference structure would fall foul of plausible constraints upon preference attribution.

Dohrn's first objection rests upon the following condition, which he thinks holds in every case of genuine agency:

> **Necessary Condition**: Doing (acting so as to make real) O_1 rather than O_2 is a case of conscious, deliberate, voluntary agency only if one does not prefer O_2 over O_1. (Dohrn, 2026, p. 235)

We know that Jean Valjean did in fact turn himself in (O_1). So, by the necessary condition, it must be the case that he did not prefer not turning himself in (O_2) to turning himself (O_1). But, if he has quasi-cyclical preferences, then he *does* prefer not turning himself in (O_2) to turning himself (O_1). So, Dohrn concludes, this cannot be a case of genuine action.

The problem here is that Necessary Condition is not convincing as it stands. On the face of it, there are obvious counter-examples. Suppose that I prefer eating chocolate ice cream (O_2) to eating raspberry sorbet (O_1). I might nonetheless act so as to make (O_2) real, if, for example, I am unaware that the result of my action will be making (O_1) real. So, at a minimum, the condition needs to be reformulated to make clear that one is doing (O_1) *qua* (O_1) – that is to say, that one is doing O_1 knowingly and under that description. Thus:

> **Necessary Condition (*)**: Doing (acting so as to make real) O_1 rather than O_2, *when one does O_1 knowingly and qua O_1*, is a case of conscious, deliberate, voluntary agency only if one does not prefer O_2 over O_1.

But now we need to ask what it is to do O_1 knowingly and *qua* O_1. In many cases, there will not be multiple ways of describing O_1 – or more accurately, because there are always multiple ways of describing any action, there will not be multiple ways of describing (framing) O_1 that have potential implications for the structure of the agent's preferences. In such cases, as stated earlier, we can to all intents and purposes make decisions as if we were making them over a purely extensional opportunity set (even though Necessary Condition (*) is not formulated extensionally).

The whole point of the Jean Valjean example, though, is that this is a case where there are multiple ways of describing (framing) O_1 that *do* have potential implications for the structure of the agent's preferences. And so, when applying Necessary Condition (*), we need to bring in one of the ways in which Jean Valjean is framing (O_1). So, let us say that what he actually does is $F_A(O_1)$ – where this denotes O_1 as framed in terms of F_A. But once we have frames in play, the argument breaks down, because there is nothing that Jean Valjean prefers to $F_A(O_1)$. It is true that there is something that Jean Valjean prefers to $F_B(O_1)$, but $F_A(O_1)$ and $F_B(O_1)$ are different objects of choice, even though (let us assume) Jean Valjean knows that they are different ways of framing the act of turning oneself in. This is the basic difference between having cyclical preferences and having quasi-cyclical preferences. Cyclical preferences fall foul of the problem that, for any possible action, there is some action that the agent prefers to it. Preferences that are quasi-cyclical do not have this problem, however, because they arise in contexts that are too complex to support the simplifying assumption that we are making choices over a purely extensional decision set.

The first objection does not, therefore, provide independent grounds for rejecting the idea of quasi-cyclical preferences, because nobody who accepts that there are preferences with a quasi-cyclical structure will be moved by Dohrn's description of the situation. This is a good moment to shift attention to Dohrn's second objection, because he offers a direct argument from basic principles that he takes to govern preference-attribution to the incoherence of what I am terming quasi-cyclicality.

Here is what Dohrn describes as a heuristic principle governing how preferences are attributed:

> An agent either prefers O_1 to O_2 or is indifferent between O_1 to O_2 and does not prefer O_2 to O_1 if the agent consciously, upon careful reflection, and voluntarily chooses O_1 over O_2 when she is placed so as to choose either O_1 or O_2 and does not ignore any fact that would influence her choice. (Dohrn, 2026, p. 235)

Note that this is a sufficiency claim, not a necessity claim. The claim is that the presence of a conscious, reflective, and voluntary choice of O_1 over O_2 entails, at a minimum, that O_2 not be preferred over O_1 – not that preferring O_1 over O_2 requires a conscious, reflective, and voluntary choice of O_1 over O_2.

Dohrn then comments:

> A criterion of this sort does not sit well with attributing quasi-cyclical preferences. Take again the example of Macbeth: as a matter of fact, Macbeth kills Duncan (O_2) rather than not killing him (O_1). Yet by assumption, he prefers both (i) Fulfilling his Double Duty (O_1) to Murdering the King (O_2) and (ii) Bravely Taking the Throne (O_2) to Backing Away from his Resolution to Make the Prophecy Come True (O_1). Only framed preferences (ii) conform to the choice. It seems plausible that we can attribute a preference for O_2 over O_1 to Macbeth. Given the transparency of framed preferences, these preferences correspond to (ii). Yet what reason do we have to attribute framed preferences (i) rather than mere wishes, desires etc.? (Dohrn, 2026, p. 235)

There is something of an equivocation running through this passage. When he applies the heuristic principle quoted above, he does so to distinguish between the following:

(i) *Fulfilling his Double Duty* (O_1) is preferred to *Murdering the King* (O_2)
(ii) *Bravely Taking the Throne* (O_2) is preferred to *Backing Away from his Resolution to Make the Prophecy Come True* (O_1).

His initial point is that what Macbeth actually does supports (maybe: requires) attributing to him the second preference, which is of course a frame-relative preference. But he then redescribes what is going on as Macbeth preferring O_2 over O_1, with the frames dropping out of the picture. Why does he make this move, which on its face seems somewhat question-begging?

The most likely explanation (to my mind) is that he was never really talking about frame-relative preferences at all. If he had been exploring the relation between preference and choice at the level of frame-relative preferences he would have acknowledged that it is far from straightforward to apply his sufficiency criterion. In situations such as Macbeth's (or Agamemnon's, or Jean Valjean's), there is no straightforward inference from an action *tout court* to a preference. What matters is not just what somebody does, but how and why they do it. For Macbeth's killing Duncan to warrant attributing to him the preferences outlined in (ii), it must be the case not just that he kills Duncan, but that he kills Duncan in the interests of *Bravely Taking the Throne* – and, *pari passu*, it must be the case, not just that he rejects the option of not killing Duncan, but that he rejects it under the frame *Backing Away from his Resolution to Make the Prophecy Come True*. Without these qualifications, there would be no reason to attribute to Macbeth the preference in (ii), rather than the inversion of the preference in (i), viz.

(i) * *Murdering the King* (O_2) is preferred to *Fulfilling his Double Duty* (O_1).

As I read the play, however, attributing to Macbeth the preference in (i)* would completely misrepresent his conative and affective state, even though it is true that he opts for O_2 over O_1.

I suspect, however, that at this point Dohrn will ask how we might go about attributing to Macbeth a preference structure richer than be straightforwardly read off from his actions. What grounds might we have, in other words, for distinguishing between the preference described in (ii) and the preference described in (i)*? We know that Macbeth does actually choose O_2 over O_1, and so, per the revised version of the Necessary Condition, we can be confident that he does prefer O_2 to O_1, provided that he chooses O_2 knowingly and *qua* O_2. But both ways of framing O_2 would satisfy this condition, and so what grounds would there be for attributing one rather than the other?

At this point it helps to remember that verbal behavior is behavior too! In the case of Macbeth we are fortunate to have a lot of verbal behavior to go on in determining how he was framing the action of killing Duncan and the evidence is, I believe, overwhelming that (ii) captures the structure of his preferences much better than (i)*. Of course, one cannot always read people's preferences straightforwardly off their verbal behavior, any more than one can read them straightforwardly off their non-verbal behavior. There are always possibilities of self-deception, insincerity,

and so forth. Attributing preferences is a holistic matter of inference to the best explanation, weighing what people do in the light of what they say (and do not say) about the different options they take themselves to have. It is not always straightforward, particularly not in the type of situation where it might be appropriate to attribute quasi-cyclical preferences. But it is an activity that we engage in all the time, because really what is at stake is identifying the *reasons* that people have for acting the way they do. It is not easy to imagine what social life would be like if we were not able to make the sort of discriminations that would allow us to attribute the preference described in (ii) rather than the preference described in (i)*.

Dohrn might respond by conceding the point with respect to preference-attributions that are grounded in actual choice, while holding out against preference-attributions that go against choice behavior. This would still be an effective strategy against the possibility of attributing quasi-cyclical preferences. This seems a tough row to hoe, however. It is not remotely plausible to say that we can *never* attribute preferences that go against choice behavior. There are many situations where we want to say that people act counter-preferentially. In fact this is often taken to be an index of practical irrationality. This is most obvious in cases of lack of self-control or weakness of will —how could we even describe the phenomenon if we were not allowed to talk about people preferring *a* over *b*, but nonetheless doing *b*? There also seem to be cases where people are manipulated into acting against what it seems right to describe as their real preferences. This is a very plausible way of describing what happens when people are seduced by cults, as well as in more quotidian types of scams. In all of these cases, when we say that *a* is preferred to *b*, even though the agent actually performs *b*, we are really saying that the agent would have performed *a* if various other factors had not been in play (self-deception, manipulation, undue influence, and so on). Plainly in cases such as those we can have perfectly good grounds for making such counterfactual claims. So what reasons could there be for denying that we can attribute preferences in situations that invite analysis in terms of quasi-cyclical preferences?

At this point, the only viable strategy for Dohrn would be to deny that these are really cases of *preference*. As he puts it in the passage quoted above, "what reason do we have to attribute framed preferences (i) rather than mere wishes, desires etc.?". As he notes, this is a strategy that has also been proposed by others (e.g., Guala, 2022) as an objection to quasi-cyclical preferences. It seems reasonable to expect, though, that it should generalize to cover the other sorts of cases discussed in the previous chapter. The problem is that it does not describe them accurately. There is a fundamental difference between (a) doing something even though I desire not to do it, and (b) doing something even though I prefer not doing it to doing it. There is nothing particularly puzzling or troubling about the first. Life is full of things that one does even though one desires something else incompatible with doing it —you might go to the gym, for example, even though at some level you desire to stay at home— or even simply desire not to go to the gym. This is hardly an example of practical irrationality. It is simply a case where one desire outweighs another. Cases of the second type, however, seem qualitatively different. The problem in cases of lack of self-control is, to put it in terms of desire, that I desire *not* to do what I actually do more than I desire to do it —as when, for example, I stay at home even though I desire to go to the gym more than I desire to stay at home. That is why these are typically described as cases of practical irrationality. The problem is not just that I have conflicting desires. It is that I act against those desires. Once the matter is put in these terms, however, we are really talking about preferences. The first case, where my desire to go to the gym outweighs my desire to stay at home, is naturally described in terms of my preferring going to the gym to staying at home, and then acting in accordance with that preference. In the second case, I have the same preference, but act against it. We cannot describe this case, or any of the others, without talking about agents having preferences that run counter to what they actually do. Since we do describe these cases all the time, it follows that we have techniques that work for attributing preferences (not just wishes and desires) that run counter to what we actually do. Those are the very techniques that we can use to attribute preferences in cases of quasi-cyclicality.

To conclude this section, Dohrn's thought-provoking arguments against the possibility of quasi-cyclical preferences are not in the last analysis convincing, despite opening up a range of very interesting questions. But quasi-cyclical preferences can be possible without being rational, and so the next section will explore Dohrn's objections to my arguments for the existence of rational quasi-cyclical preferences.

5.3. Preferences that are quasi-cyclical and rational?

In section 3 of his paper, Dohrn suspends his skepticism about quasi-cyclicality and assesses my arguments that there can be *rational* quasi-cyclical preferences. That such patterns of preference exist is one of the main claims of *Frame It Again: New Tools for Rational Decision-Making*, where quasi-cyclical preferences are the new tools referred to in the subtitle. I try to make the case in multiple domains, but the basic line of argument is similar in each case, summarized in Bermúdez (2022) in the following three propositions:

> (H1):
>
> Frames and framing factor into decision-making by making one dimension/attribute/ value of the decision problem highly salient, which influences how the subject engages emotionally and affectively.
>
> (H2):
>
> Quasi-cyclical preferences are likely to be found in decision-problems that are sufficiently complex and multi-faceted that they cannot be subsumed under a single dimension/attribute/value. Different frames engage different affective and emotional responses, which the decision-maker cannot resolve either by ignoring all the frames except one or by subsuming them into a larger frame.
>
> (H3):
>
> Framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faced response to a complex and multi-faceted situation.

The approach here is diametrically opposed to the standard emphasis on small-world decision-making, focusing on the hard cases rather than the easy ones. It may be that hard cases make bad law, but they do, I claim, provide important insights into rational decision-making.

Dohrn's criticisms arise out of an independently interesting distinction that he draws between two different ways of evaluating the rationality of a set of preferences. We can assess the rationality of the decision-making processes that lead up to those preferences. This is what he calls *input-rationality*. Alternatively, we can consider the *output-rationality* of a set of preferences, by assessing how those preferences serve to guide action. Dohrn applies the distinction to his example of Jean Valjean, as discussed in the previous section.

Dohrn concedes (somewhat guardedly - see his n. 8) that Jean Valjean's quasi-cyclical preferences could be described as input-rational, thereby accepting my suggestion in (H3): " In this sense, one may argue, following Bermúdez, that Jean Valjean's practice of framing and reframing is an adequate way of responding to his morally complex situation, and this practice may rationalize forming not only conflicting desires or wishes in responding to conflicting demands, but also forming quasi-cyclical preferences." (Dohrn, 2026, p. 237) However, drawing on his arguments (discussed in the previous section) about the connection between preference and action, he thinks that it is in principle impossible for quasi-cyclical preferences to be output-rational. This is because, he claims, whatever option an agent with quasi-cyclical preferences decides upon, the quasi-cyclicality of his preferences ensures that there will be another option that he prefers to it. Here is the argument:

> [...] whenever one has quasi-cyclical preferences $FA(O_1)$ over $FB(O_2)$ and $FC(O_2)$ over $FD(O_1)$, one cannot choose one action without violating the demands of rationality: if one chooses O_1, there is an alternative O_2 that one prefers to it (albeit under a certain frame), and if one chooses O_2, there is an alternative O_1 that one prefers to it (albeit under another frame), and one knows this. In sum, by Bermúdez's own lights, quasi-cyclical preferences are just as irrational as cyclical preferences as far as output-rationality is concerned. (Dohrn, 2026, p. 238)

if Dohrn's analysis is correct, then quasi-cyclical preferences cannot guide rational action, since one cannot be acting rationally in performing, say, O_1 if there is an (available) option that one prefers to O_1.

Of course, as Dohrn recognizes in his two parentheses, his characterization of the preference structure is somewhat contentious, in ways that were discussed in the previous section. Quasi-rational preferences only block rational agency (as he puts it) if one ignores the frames. After all, as Dorhn himself describes the situation, there is nothing that Jean Valjean prefers to $FA(O_1)$. There is something that he prefers to $FC(O_1)$, but the whole point of the analysis is that $FA(O_1) \neq FC(O_1)$. So, on this point, it certainly seems open to me to dig my heels in and insist that in this case we are dealing with frame-relative preferences. From a dialectical point of view, it should be noted that it seems difficult for Dohrn to hold simultaneously (a) that Jean Valjean's frame-relative preferences are input-rational, and (b) that we should ignore the frames when we are considering output-rationality.

Be that as it may, there is a real difficulty not too far away, noted by Dohrn and earlier by Richard Pettigrew (2022). There is an inherent instability about quasi-rational preferences, in the sense that what agents choose depends upon which frame they are in and *ex hypothesi* these are situations where the dominant framing is constantly changing. So, what does happen when someone with quasi-cyclical preferences acts? How does an essentially unstable preference structure get resolved in a way that allows it to be translated into action? It looks as if there are two alternatives, neither particularly palatable from the perspective of someone who believes in the rationality of quasi-cyclical preferences. The first alternative, proposed by Pettigrew, is that the process of resolution is one of settling on a set of non-cyclical preferences (most obviously, by settling on one frame, to the complete exclusion of the other). If this is the correct account, then it is hard to make the case that quasi-cyclical preferences are rational, since (rational) action is only made by possible by eliminating quasi-cyclicality. The second alternative, to which Dohrn seems sympathetic, is that agents with quasi-cyclical preferences simply ending up acting on the basis of whichever frame happens to be salient at the time of action. If this is the correct account, then, as Dohrn puts it, "acting on quasi-cyclical preferences becomes subject to accidental circumstances in a way that conflict with our expectancies for rational action. It just so happens that one framing rather than the other is salient so as to guide one's action." (Dohrn, 2026, p. 238)

The first option is no doubt a correct description of *some* cases. Many agents with quasi-cyclical preferences will strive to restore consistency (in the textbook sense). In *Frame It Again* I pointed out that there are constraints both upon how it can be rational to frame things and upon the rational co-tenability of different frames (see particularly Ch. 10). It may well be that thinking through those constraints in a particular case will lead a decision-maker to modulate from quasi-cyclical preferences to preferences that are completely non-cyclical. Or perhaps they will find an overarching frame that resolves quasi-cyclicality. It would be foolish to deny that these things can take place, but no less undesirable (I think) to insist either that they always take place as a matter of fact, or that rational decision-making always requires them as a matter of principle. So, we should look at the second option. Here, Dohrn is surely correct that, in some sense of 'salience', an agent who acts against a background of quasi-cyclical preferences will act on the reasons provided by whatever frame is salient at the time of action. But does that make action arbitrary in the way he implies? Everything depends, of course, on how and why a frame becomes salient.

There certainly seem to be cases of non-arbitrary salience. Self-control is a good example (as discussed in Chapter 7 of *Frame It Again*, and earlier in Bermúdez, 2018a). There is a longstanding puzzle of explaining how the exercise of self-control is even possible within the constraints of traditional theories of choice, if we take self-control to involve overriding short-term satisfaction at the point of choice in the interests of longer-term gain, and we take it as axiomatic that agents will do what they most prefer. If longer-term gain is desired more strongly than short-term satisfaction, then there does not appear to be a problem at all —it is an easy win for long-term gain. But if, on the other hand, the desire for short-term motivation is stronger, then it is hard to see how longer-term gain can get a foothold at all. My suggestion is that we can do justice by taking into account different ways of framing the short-term satisfaction and the longer-term gain in a way that gives the agent quasi-cyclical preferences. The agent prefers *Immediate Satisfaction* to *Longer-term Gain*. That is why there is a problem. On the other hand, the agent also prefers *Resisting Temptation* to

Immediate Satisfaction. These are quasi-cyclical preferences, because *Longer-term Gain* and *Resisting Temptation* are the same option, differently framed, and the agent exercising self-control surely knows this. For self-control actually to be exercised, the *Resisting Temptation* frame must be more salient than the *Longer-term Gain* frame. But in many cases it would be odd to describe this as arbitrary, and hence irrational. People who exercise self-control often do so as part of a carefully thought-out life-plan and as the result of painstakingly developed techniques for avoiding temptation, and so on. In many cases that will surely make it reasonable to describe their self-controlled action as rational.

Extrapolating, quasi-cyclical preferences do indeed present a challenge when we think about the rationality of action, because determining whether an action is rational or not will depend upon a broader assessment of the rationality of a particular frame being salient. That broader assessment, in turn, will need to go beyond the immediate context of the action, taking a much more holistic perspective on the agent's motivations and psychological profile. The lesson to draw is that, in the type of cases that can plausibly be analyzed in terms of quasi-cyclical preferences, the rationality of an action cannot simply be read off the structure of an agent's preferences. But that seems the right result for complex situations.

Still, one might reasonably ask what can be said about first-person reasoning with frames, as opposed to the third person determinations of whether an action is rational or not.[25] It would be natural to ask, for example, how and when, in situations where an agent has quasi-cyclical preferences, the rationality of an action can be a function of how an agent reasons with frames —and, more generally, of how an agent can reason within and across frames. This brings us to Sarah Fisher's very interesting contribution to this special issue, which will be the focus of the next section.

5.4. Tackling discursive deadlock

The previous section briefly considered an important question that arises for anyone who thinks, as I do, that quasi-cyclical preferences can be rational. Quasi-cyclical preferences emerge when an agent is knowingly framing a single option, or multiple options, in different ways. In §5.1 we looked at various ways in which such multiple framing might come about, as a function, *inter alia*, of different forms of emotional engagement and perceived similarities. Frames do not come out of nowhere. They are reflections (and refractions) of an agent's broader psychological and cognitive profile. But, one might ask, to what extent are they under an agent's control? And, in particular, to what extent can agents reason within and across frames? These are pressing questions that arise even if it is conceded (with (H3) above) that framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faced response to a complex and multi-faceted situation. That simply tells us the circumstances in which it can be rationally permissible to have quasi-cyclical preferences, but not how an agent might reason their way *to* quasi-cyclical preferences or how they might reason *with* them.

Both of these questions are tackled in Chs. 10 and 11 of *Frame It Again*, where I offered (a) an account of some of the normative constraints that I take to hold upon an agent holding multiple frames simultaneously, and (b) a model of what I termed *frame-sensitive* or *non-Archimedean* reasoning. I argued that (a) and (b) could be deployed to help defuse the *discursive deadlock* that has become endemic in political and other forms of social discourse. Sarah Fisher's original and insightful contribution takes issue with my suggestions for using non-Archimedean reasoning to defuse discursive deadlock, because she thinks that I focus too heavily on how frames provide reasons, neglecting the ways in which frames serve to signal identity and affiliation. Her (somewhat dispiriting) conclusion is that the effect of deploying frames can be the very opposite of what I propose, namely, to reinforce the very types of communicative dynamic that generate tribalism and polarization.

---

[25] For more on this distinction between thinking about rationality in first- and third-person terms, see my *Decision Theory and Rationality* (Bermúdez, 2009), particularly Ch. 1.

To take a step back, here is a passage from *Frame It Again* that Fisher also quotes to illustrate the basic phenomenon of discursive deadlock:

> Suppose that I am trying to make my mind up on a complicated and emotive issue, such as whether to support a political candidate whose views are broadly aligned with mine, except on the issue of the pledge of allegiance. He believes, while I do not, that children should have the opportunity to pledge allegiance to the flag of the United States every morning at the start of the school day. [Footnote omitted] Underlying the dispute is our framing the pledge in different ways. For him, it is an expression of patriotism and a reverence for the values that he thinks underpin the American constitution and the American way of life. In contrast, I, like many people who grew up in Europe, have an inherent distrust of flags, which I tend to frame from a historical perspective. Hitler, Mussolini, and Franco all came wrapped up in flags and the dominant image of a flag to my mind is of hundreds of thousands of people saluting the swastika at the Nuremberg rallies. (Bermúdez, 2020, p. 235)

I term the two frames here Flag and No Flag.

The graphic in Figure 1 presents the four elements of the model of frame-sensitive reasoning with which Fisher engages.

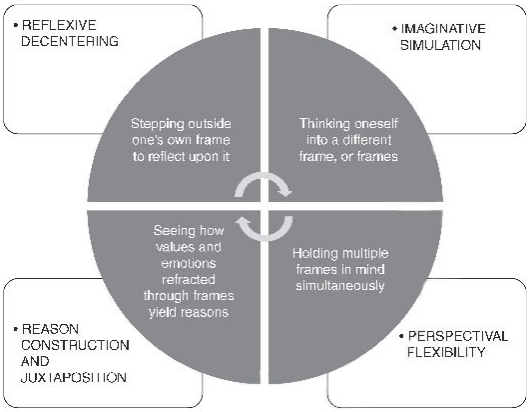

Figure 1

Figure 11.1 from Bermúdez (2020)

In a little more detail:

> *Reflexive decentering*. Flag and No Flag need to recognize that they are each framing the pledge of allegiance issue in fundamentally different ways, and that these framing differences are the source of their disagreement and deadlock. So, a frame-sensitive reasoner confronted with a sufficiently complex and multifaceted decision problem will need to start by stepping outside their initial framing of an action or outcome in order to reflect upon the frame itself.
>
> *Imaginative simulation*. Frame-sensitive reasoners then need to imagine what it would be like to frame things completely differently, and then to simulate actually being in that frame. This simulation extends to the beliefs, desires, and other affective attitudes of someone in that other frame, so that the agent sees the world from the perspective of that frame.
>
> *Perspectival flexibility*. Frame-sensitive reasoning involves considering multiple frames simultaneously. Rational preferences are held for reasons and different frames bring different reasons into play. One needs to be able to adopt multiple frames simultaneously in order to evaluate how those reasons interact, and hence fully to appreciate the complexity of the decision problem.

> *Reason construction and juxtaposition*. Different frames bring different reasons into play in multiple ways. How an outcome or action is framed can foreground some reason-giving feature while downplaying another – as when Flag foregrounds one sentence in the pledge of allegiance while No Flag foregrounds a different one. Another way is for the framing to bring into play a reason-giving similarity to some other (appropriately framed) action or outcome. Or framing an outcome a certain way can express a particular value in a particular way. Appreciating these requires being able to see how values and emotions refracted through frames can yield reasons, and how different frames can yield different reasons.

As Fisher points out, this all depends on thinking of frames primarily in terms of the reasons that they provide. The exercise is one of appreciating how different frames can provide different reasons and thereby coming to a fuller understanding of what is, *ex hypothesi*, a complex and multi-faceted decision problem. One possibility (the ideal case, no doubt) is that this process will reveal hitherto neglected common ground. Even if it fails to do that, one might hope for a decrease in animosity, and correlatively a move towards civil discourse.

Fisher finds these aims laudable, but overly optimistic, taking issue primarily with my focus on how frames operate to provide reasons. While I see frames as playing important roles in processes of generating, discovering, and articulating reasons, and hence more broadly as powerful tools in rational debate and conflict resolution, she argues that frames can work in ways that run counter to, and serve to undermine, those reason-governed activities. To my rationalistic way of looking at frames, she contrasts a view of frames that is, if not anti-rationalistic, then a-rationalistic.

To appreciate her argument we need to start with how she understands frames:

> Considered in the broadest sense, *framing* and *frames* concern how things are presented—whether privately in individual reflection or publicly in social interactions. Our narrower focus here will be on how entities, individuals, or states of affairs are *articulated in language*.
>
> Here I am cleaving closely to how 'framing' and 'frame' are used by Tversky and Kahneman in their seminal studies of framing effects (Tversky & Kahneman, 1981, 1986; Kahneman & Tversky, 1984) and by many subsequent psychologists (see, for example, the survey of framing studies by Levin *et al.* (1998)). These researchers typically characterise the frames they are investigating as alternative *descriptions*.
>
> It should be noted that the psychological research focuses on articulations assumed to be logically equivalent (such as 'one quarter full' and 'three-quarters empty', or 'eighty percent die' and 'twenty percent survive'). Following Bermúdez (2020), I will use 'frame' more broadly here, to include articulations that are not logically equivalent but are nevertheless understood to be concern the same thing in the context at hand. For example, I might characterise checking my emails as "my least favourite task" or "the next item on my to-do list". Insofar as these two sequences of words are understood to be characterising one and the same activity, they constitute alternative frames for it. (Fisher, 2026, p. 246)

This highly language-focused way of thinking about frames is very much in line with influential work in cognitive linguistics (e.g., Lakoff and Johnson, 1980; Lakoff, 2004), but her view of framing is considerably narrower than mine, and this will turn out to be important.

In thinking about how frames function lexically, Fisher sees an important analogy with the pragmatic function of slurs and dogwhistles, both of which she interprets, not implausibly, as functioning as linguistic tools for signaling identity and affiliation. She summarizes her overall position:

> I wish to suggest that alternative framings of complex social and political issues, of the kind Bermúdez discusses, can themselves signal affiliation with distinct communities of speakers. Indeed, as will be discussed, someone's choice of frame in public debate and collective decision-making might ultimately have less to do with the reason elucidated by the frame and more to do with building and

> maintaining social alliances. By the same token, the adoption of an opponent's frame could function less as an exercise in rational due diligence than as the enactment of an antagonistic social identity, with concomitant risks of alienation and ostracization. (Fisher, 2026, p. 251)

If frames function primarily as social signals in this way, then the prospects for the type of non-Archimedean reasoning sketched out above start to look rather dim. On Fisher's view, the frames are not the problem, but rather an effect of the problem: "On this analysis, it is upstream socialisation and identity formation which may often turn out to be the real source of discursive deadlock, not frame-sensitive appreciation of only some reasons" (2026, p. 253). If this is right, then attempts to bridge over to different framings are likely to be not only very difficult and perhaps traumatic, but also fraught with social risk.

One of the most interesting parts of Fisher's argument is her analysis of how social media work to reinforce the signaling function of frames and, correlatively, to crowd out the reason-giving and deliberative aspects of frames. She highlights three structural features of social media in particular.

> *Connectivity* Social media platforms are designed to promote social bonding – with friends, family, and total strangers, rather than reasoned debate or transmitting professionally produced content. The goal of social bonding promotes signaling of shared identities and affiliations and so in turn promotes ways of linguistic framing that achieve that signaling (the in-group frames) – and correlatively of discouraging adoption of frames that do not belong within the group.
>
> *Publicity and social sanctions* The amplifying effect of social media is well-known, as its capacity for generating vituperative backlash. This can come from in-groups no less than out-groups - in the case of in-groups it can be in the service of policing shared social norms. The fear of in-group policing is a further disincentive to publicly adopting any sort of out-group framing.
>
> *Decontextualization* Social media interventions are typically short and their appearance in individual feeds is algorithmically determined in ways that obscure whatever context they might originally have had, and clarifications in comments easily get buried. Given the potential social sanctions noted above, this has the effect of promoting discourse in which identity and affiliation are clearly signaled, to avoid the costs of ambiguity and misinterpretation. (Fisher, 2026, 248-249)

In sum, the communicative dynamics of social media platforms reinforce discursive deadlock through structures that encourage in-group framing and actively work against what I call reflexive decentering and perspectival flexibility, without which non-Archimedean reasoning cannot get started.

What Fisher says about the pernicious effects of social media is insightful, and she has promising suggestions for mitigating those effects, at least partially. I am not yet convinced, though, that the prospects are as dim as she takes them to be for using frame-sensitive reasoning to break discursive deadlock. The basic problem with her argument is that it mislocates the terrain, so to speak, of non-Archimedean reasoning, for reasons that go back to her original characterization of what frames are and how they work. Let us look again at her diagnosis of discursive deadlock: " . . . it is upstream socialisation and identity formation which may often turn out to be the real source of discursive deadlock, not frame-sensitive appreciation of only some reasons" (2026, p. 253). I do not think that we should accept this distinction. One of the basic take-home messages of *Frame It Again* and my other works on framing is that the type of frames in which I am most interested are key elements in what Fisher terms "upstream socialisation and identity formation" (Fisher, 2026, p. 253). Of course, these are not the only frames that there are, but it is important to distinguish the forms of framing that I am discussing from those that are studied by cognitive linguists and experimental psychologists.

In a recent paper (Bermúdez, 2025), I work with Erving Goffman's very general notion of frames as "schemata of interpretation" that allow individuals "to locate, perceive, identify, and label" (1974, p. 21). Frames make salient one dimension, attribute, or value of a situation or decision problem. But this can occur in different ways. The classic framing experiments, replicated many times since Tverksy and Kahneman (1981), show that people tend to be risk-averse when

outcomes are framed in terms of potential gain and risk-seeking when they are framed in terms of potential losses.[26] We can think about it in terms of priming. The gain frame priming risk-aversion and the loss frame priming risk-seeking.

In the social media contexts that Fisher discusses, frames seem to be functioning in something like this priming sort of way. They signal affiliation and identity by eliciting, perhaps even provoking, certain types of response, which is part of the explanation for the visceral reactions that Fisher describes. There is a type of triggering effect at work. That this is an important phenomenon is beyond dispute. And Fisher is surely correct in two of her main claims. First, the priming/triggering effect of this type of framing is primarily a linguistic phenomenon (and, in fact, a linguistic phenomenon that sometimes co-opts the slurs, dogwhistles, and related phenomena that she discusses in the first part of her paper). And second, *these* types of framing are essentially upstream of identity formation and socialisation. This is a discursive mode into which one enters with one's identity fully formed, so to speak.

There are obviously many differences between how frames operate to prime, say, risk-seeking behavior, on the one hand, and, say, xenophobia, on the other. For one thing, the first is typically an experimental artefact (some would say: the result of experimental manipulation), while the second can be consciously adopted and refined. But the basic commonality is that they are both essentially non-reflective. By this I mean, not that no thought can go into them, but rather that they are not part of an agent's deliberative reflection on how and why they should act, what they should believe, or what they should want. They are not, in short, part of what John McDowell calls "the space of reasons".

In this regard, then, I take them to be fundamentally different from the types of frames on which I focus in *Frame It Again* and elsewhere. The features of a situation, thing, or action that these reflective frames make salient are precisely their reason-giving features. Of course, here as elsewhere, there are degrees. Shakespearean soliloquies capture the most rarified and elevated forms of deliberative reflection. Likewise for the other literary examples. Other forms of framing are much more quotidian. The self-control example is considerably less dramatic than Macbeth. But it is still in the space of reasons. My suggestion is that an agent grappling with temptation, whether on a one-off basis or when trying to develop an overarching strategy, can use different ways of framing the short-term and long-term gains as part of the decision-making process.

The same holds for the game-theoretic examples discussed in Ch. 9 of *Frame It Again* (and briefly discussed in Dohrn's contribution). I took inspiration from Michael Bacharach's model of what he calls mode-P reasoning, which adopts a "We"- frame, rather than an individualistic "I"-frame (Bacharach, 2006). One of my points of difference with Bacharach, however, concerns how the shift from the "I"-frame to the "We"-frame occurs. Whereas Bacharach explicitly sees this as a priming effect (where the shift to the "We"-frame is primed by structural features of the social interaction = game) I argue at length that adopting the "We"-frame is something that can, and should, be done for reasons – not least because there are many occasions (including the very situations of tribalism and group-think against which Fisher so eloquently warns us) on which one might hope that reason would direct us away from the "We"-frame.

In sum, while I find Fisher's discussion of frames both insightful and enlightening, and her cautionary comments about framing in social media well taken, what she says is not *directly* relevant to the type of framing on which I place most emphasis. The frames that she discusses all fall outside of the space of reasons and are, in her phrase, "upstream" of processes of socialization and identity formation, whereas what I want to emphasize is the reason-governed role that frames can play in those very processes.

* * *

My thanks once again to Daniel Dohrn and Sarah Fisher for their stimulating and complementary discussions. I have greatly benefited from the opportunity to think through the important issues that they raise.

[26] See Chs. 1 and 2 of Bermúdez (2020) for illustrations and references from multiple domains.

6. Active inference, the free energy principle, and Marr's tri-level hypothesis: Reply to Jurjako

In *Philosophy of Psychology: A Contemporary Introduction* (Bermúdez, 2005b) and the first two editions of my textbook *Cognitive Science: An Introduction to the Science of the Mind* (Bermúdez, 2010, 2014) I made some fairly strong claims about the limitations as a general tool for cognitive science of the type of tri-level analysis pioneered by David Marr (Marr, 1982). Marr's distinction between computational, algorithmic, and implementational levels, his discussion of the relations between them, and his use of them to give a detailed analysis of the visual system have rightly been thought to be paradigms of cognitive science and one of the discipline's real success stories. But, I argued, there are in principle limits to how widely Marr's tri-level hypothesis can be applied in cognitive science. I conjectured that it is probably applicable only to modular systems and was very definite that it could not be applied to the mind as a whole.

As Jurjako (2026) tactfully points out, that was a while ago and things have moved on since I made those arguments. He offers an intriguing argument that more recent work on active inference and the free energy principle has opened up possibilities for applying Marr's tri-level model much more generally. In fact, he makes the strong claim that the free-energy principle can be used to give a computational account of the function of the mind as a whole, in a way that allows us to model it algorithmically and then explore how its algorithmic basis is neurally implemented.

Jurjako's paper raises a number of interesting and important issues, and I am grateful to him for pointing out the importance of the active inference paradigm and the free energy principle, both in connection with Marr's tri-level hypothesis and more generally in cognitive science. This response, though, explains why I think that, despite the theory's many theoretical merits and explanatory power, analyses of cognition based on the free energy principle do not fit the Marrian template. The first section briefly recapitulates my reasons for skepticism about extending Marr's tri-level model beyond the realm of the modular. The second section presents Jurjako's principal claims, to which I respond in the third and final section.

6.1. Marr's tri-level hypothesis

Marr famously suggested that certain types of cognitive systems, such as the early visual system, can be analyzed at three different levels.

> The computational level
>
> Analysis at the computational level specifies the system's function or role by specifying the information-processing task that the system is configured to solve. Marr's computational analysis of the early visual system is, in essence, that its role is to transform information from the retina into a representation of the three-dimensional shape and spatial arrangement of an object.
>
> The algorithmic level
>
> An algorithmic analysis specifies algorithms that perform the information-processing task identified at the computational level. Information-processing algorithms are step-by-step procedures for solving information-processing problems, where an algorithm is a finite set of instructions that are mechanical and automatic (a computer program, for example).
>
> The implementational level
>
> Analysis at the implementational level investigates how the algorithms specified at the algorithmic level are neurally realized, where this involves finding the neural bases of the relevant inputs and outputs and identifying neural mechanisms that will transform the former into the latter. Here we are squarely in the domain of neuroscience (which, of course, itself operates at multiple different levels of analysis).

In the posthumously published *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Marr 1982), Marr applied this tri-level model to the human visual system. One of the basic principles of Marr's analysis of the visual system is that it is *modular* —that is to say, the overarching task of the visual system (which is to compute representations of the distal environment on the basis of changes in light intensity at the retina) is achieved through a series of discrete sub-systems that work quickly to provide rapid solutions to highly specific problems. These modular sub-systems are *domain specific* (carrying out a very circumscribed job with a fixed field of application), *informationally encapsulated* (i.e., not sensitive to background knowledge and expectations), and *fast* (transforming input to output quickly enough to be used in the on-line control of action.

There are two reasons, I argued, for thinking that a Marr-style analysis will (probably) only work for modular systems. The first has to do with task analysis at the computational level, which requires the system to be performing functions that are sufficiently circumscribed and determinate to be algorithmically carried out. Only when we are dealing with systems that are domain-specific and encapsulated, I suggested, will we be able to find tasks that are specific enough and narrow enough to be *algorithmically computed*. The second reason for thinking that Marr's tri-level approach works best (and perhaps only) for modular systems is that the algorithms specified at the algorithmic level must be computationally tractable. They can only contain a limited number of representational primitives and possible parameters of variation.

Informational encapsulation is a good way (the only way?) to secure computational tractability, since it ensures that the algorithm will be running on a limited range of inputs. Non-modular processing runs very quickly into versions of the so-called *frame problem* —the problem of identifying, in any problem-situation, all and only the information that is relevant to solving the problem without explicitly representing that information (McCarthy and Hayes 1969, Dennett 1987).

In any event, even if one leaves open the possibility that there might be non-modular cognitive systems for which a Marr-style analysis is appropriate, for both of these two reasons there would seem to be no prospect of applying the tri-level hypothesis to the mind as a whole. Or so I maintained. Jurjako disagrees and we turn to his arguments in the next section.

6.2. Jurjako on active inference and the free energy principle

Jurjako is a powerful advocate for the active inference approach to cognition pioneered by Karl Friston and his collaborators, the guiding idea of which is the free energy principle. I will not repeat the general introduction to the active inference and the free energy principle in section 5 of Jurjako's paper, and we will drill down into some of the details shortly. For now, all we need on the table is the basic idea that, as Jurjako puts it, "self-organizing systems maintain their integrity and resist dissipation by minimizing atypical or surprising events in their environment" (2026, p. 270). This is the 'active' part of active inference. So, let us turn to the conclusions that he draws from the existence of an active and successful research paradigm motivated by this basic principle. I particularly want to highlight two claims that he makes.

The first is well taken. Jurjako maintains, and I think he is right, that thinking about cognition and action in terms of active inference can provide a bridge between personal-level and subpersonal-level explanation. He writes:

> By providing a comprehensive and unified perspective on the mind, the FEP also offers conceptual tools to contemplate the connections between personal and subpersonal levels. As mentioned before, the FEP casts perceptual, cognitive, and behavioral abilities as different ways of accomplishing the same thing, namely minimizing free energy. In this regard, higher-level mental states and processes (including beliefs, desires, intentions, etc.) can be construed as different aspects of ways in which organisms perceive, act, and maintain homeostasis by minimizing free energy (Friston *et al.*, 2017a). Moreover, applying the FEP to the mind as a whole suggests a deep

> continuity between personal and subpersonal levels. By framing higher-level mental states as manifestations of the same fundamental drive to minimize free energy, the FEP suggests that our conscious experiences, and other mental states as captured by folk-psychology, emerge from the very same principles that govern subpersonal processes (Smith *et al.*, 2022).
>
> Thus, the active inference framework, in a sense, blurs the categorical distinction between personal and subpersonal levels by portraying them as distinct means of accomplishing the same computational tasks that enable organisms to survive and maintain homeostasis. (Jurjako, 2026, p. 273)

This seems right, and offers a promising way of thinking about what I have termed the interface problem.

Less promising, however (at least to my mind), is Jurjako's attempt to see the active inference paradigm as an application of a broadly Marrian approach. He maintains, in effect, that the free energy principle answers the problem about task analysis raised in the previous section, because we can view the function of the mind simply as the minimization of free energy. He writes:

> At the computational level, the FEP provides a general view according to which the function of the mind is minimization of free energy. Indeed, at the most general level, FEP is typically understood as a normative framework that captures "what living organisms must do to face their fundamental existential challenges (minimize their free energy) and why (to vicariously minimize the surprise of their sensory observations)" (Parr *et al.*, 2022, p. 8,). (Jurjako, 2026, p. 273)

It is certainly true, if the active inference paradigm is indeed a sound approach, that there is a level of description on which this account of the function of the mind is correct —just as there is a level of description on which the function of the body is homeostasis. But Jurjako is explicitly making a much stronger claim, namely, that this description is sufficiently precise to allow for algorithmic determination:

> At the algorithmic level of analysis, this framework provides insight by offering process level theories of cognitive phenomena (Friston *et al.*, 2017a). Process theories encompass explanatory frameworks that describe more abstract mechanistic components underlying belief updating in the brain, as well as its broader impact on an organism's interactions with the environment (see Parr *et al.*, 2022, p. viii). (Jurjako, 2026, p. 274)

This is the key claim that will occupy us for the remainder of this paper. Marr's tri-level hypothesis requires not simply that there be three levels of analysis. His model is top-down. The task analysis spells out the success conditions for the algorithm. It is not enough to pick out in very general terms an overarching task that the mind can be said to perform. The task has to be sufficiently precise and well-defined to be algorithmically computable. And, moreover, it has to be algorithmically computable in a way that is defined over representational primitives and operations that are plausible candidates for neural realization. The next section explains my doubts about either of these requirements being satisfied either by the active inference approach in general, or by the specific examples that Jurjako gives.

6.3. Active inference and algorithms

Let me begin with Jurjako's thought-provoking claim that we can provide a task analysis for the mind as a whole in terms of the minimization of free energy. As I commented in one of the passages to which Jurjako refers in his paper, (almost) everything can be given a task analysis if we are prepared to zoom out sufficiently far. We need to be careful, though, that we do not end up with a task analysis formulated at such a level of generality that it is impossible to operationalize it algorithmically. The tasks identified in the task analysis need to be specific enough to generate algorithms. It is far from clear that a task analysis that simply appeals to the free energy principle can do this.

In a recent overview paper in *Physics Reports*, Friston and colleagues give an account of the basic idea behind the general appeal to the free energy principle that is very relevant here.

> Many theories in the biological sciences are answers to the question: "what must things do, in order to exist?" The FEP turns this question on its head and asks: "if things exist, what must they do?" More formally, if we can define what it means to be something, can we identify the physics or dynamics that a thing must possess? To answer this question, the FEP calls on some mathematical truisms that follow from each other. Much like Hamilton's principle of least action,1 it is not a falsifiable theory about the way 'things' behave —it is a general description of 'things' that are defined in a particular way. As such, the FEP is not falsifiable as a mathematical statement, but its postulates may be falsifiable to the extent that they refer to a specific class of empirical phenomena that the principle aims to describe.
>
> Is such a description useful? In itself, the answer is probably no —in the sense that the principle of least action does not tell you how to throw a ball. However, the principle of least action furnishes everything we need to know to simulate the trajectory of a ball in a particular instance. In the same sense, the FEP allows one to simulate and predict the sentient behaviour of a particle, person, artefact or agent (i.e., some 'thing'). This allows one to build sentient artefacts or use simulations as observation models of particles (or people). These simulations rest upon specifying a generative model that is apt to describe the behaviour of the particle (or person) at hand. At this point, committing to a specific generative model can be taken as a commitment to a specific —and falsifiable— theory. (Friston *et al.*, 2023, p. 2)

This is very telling. Friston explicitly states that the free energy principle, considered in isolation, is not sufficiently determinate to be empirically falsifiable. It is instead, as Friston puts it, a very general characterization of the physics or dynamics that things must possess. That is to say, the free energy principle only sets the most general parameters for the class of more specific accounts that we give of particular types of activities or processes. It is only with these more specific accounts, which Friston terms generative models, that we arrive at falsifiable claims.

It is quite plain that Marr himself took task analysis at the computational level to involve substantive empirical claims. His own task analysis of the visual system was based on a hypothesis about the function of the visual system based upon well-documented patterns of breakdown in neuropsychological patients. If we assume, as Marr himself appears to have done, that a satisfactory task analysis for Marrian purposes must minimally be falsifiable, it looks as if even the originator of the active inference approach would have doubts about taking the free energy principle to provide a task analysis for the mind as a whole. Let us return, then, to a crucial passage from Jurjako's paper (at p. 274):

> At the algorithmic level of analysis, this framework provides insight by offering process level theories of cognitive phenomena (Friston *et al.*, 2017a). Process theories encompass explanatory frameworks that describe more abstract mechanistic components underlying belief updating in the brain, as well as its broader impact on an organism's interactions with the environment (see Parr *et al.*, 2022, p. viii).

If I am reading him correctly, Jurjako is suggesting that the generative models that Friston refers to in the passage just quoted should be viewed at the algorithmic level of analysis – as offering algorithms carrying out the more general task of minimizing free energy. But the passage from Friston suggests a different way of thinking about them – not as algorithms executing the task of minimizing free energy, but rather as models that share the common feature that, at a suitably general level of description, they can all be seen as minimizing free energy.

There is a more significant point in the vicinity here. The active inference approach is, methodologically speaking, grounded more fundamentally in physics than in cognitive science. As applied by Friston, it may indeed best be viewed as an instance of the dynamical systems approach to cognition, because the theoretical basis for the active inference approach is squarely grounded in the physics of dynamical systems. This comes across very clearly in Friston *et al.* (2014),

which is revealingly entitled 'Cognitive dynamics: From attractors to active inference'. This paper sets out to provide a justification of the active inference approach from the first principles of statistical dynamics. Here is how they summarize the general contours of their argument:

> Our premise is that biological self-organization is (almost) inevitable and manifests as a form of active Bayesian inference. We have previously suggested [11] that the events "within the spatial boundary of a living organism" [6] may arise from the very existence of a boundary or blanket, and that a Markov blanket may be inevitable under local coupling among dynamical systems. Here, we extend these arguments to provide a seamless progression from the basic behavior of random dynamical systems to formal (normative) accounts of the embodied or active inference that underlies cognition. To do this, we formulate flows in random dynamical systems in generalized coordinates of motion, relate this formulation to (filtering) procedures found in statistics and control theory, and then consider what this (neuronal) filtering would look like in the brain. (p. 427)

This is not the place to go deeply into the philosophy of science, but hopefully it will not be too controversial that there is a fundamental difference between the types of equation that we might find in statistical mechanics, or really anywhere else in physics, and the algorithms that are needed for Marr-type models in cognitive science. Painting with a very broad brush, physical equations set out (under certain idealizing assumptions) the relations between specific physical quantities —the relation between energy and mass in Einstein's famous equation $E = mc^2$, for example. Dynamical equations set out how specific quantities vary as a function of each other. An algorithm, in contrast, has to be a step-by-step process for computing or calculating the solution to a specific problem. Algorithms are tools for computation, not descriptions of how quantities vary as a function of each other. And there is a significant gap between writing down an equation, however descriptively accurate and/or theoretically fecund it might be, and specifying an algorithm, although, of course, an algorithm might involve writing an equation, as in the much-discussed Marr-Hildreth algorithm for edge detection (on which see more below).

To explore this point further, let us consider an equation on which Jurjako places a considerable amount of weight:

$$G(\pi) = -E_{Q(\tilde{h},\tilde{s}|\pi)}[DKL[Q(\tilde{h}|\tilde{s},\pi)||Q(\tilde{h}|\pi)]] - E_{Q(\tilde{s}|\pi)}[\log P(\tilde{s}|C)]$$

Jurjako offers a nice gloss of this equation:

> Here π denotes a policy, i.e. a sequence of actions an agent can choose. $\tilde{s}$ denotes a sequence of sensory observations that is received by an agent over time, while C denotes a set of parameters that encode the agent's preferences about various possible outcomes. *hP* denotes a sequence of hypothesized external states that evolve over time and produce or will produce the sequence of observations. In the first term, $EQ(hP,\tilde{s}|\pi)$ indicates the expected value of a distribution, where the distribution $Q(\tilde{h}|\tilde{s},\pi)$ is used for evaluating the value of the distribution $DKL[Q(\tilde{h}|\tilde{s},\pi)||Q(hP|\pi)]$, while in the second term, $EQ(\tilde{s}|\pi)$ indicates the expected value of a distribution $Q(\tilde{s}|\pi)$ that is used for evaluating $P(\tilde{s}|C)$, i.e. how expected are observations $\tilde{s}$ given a set of preferences C. As before Q denotes a probability distribution that approximates the true posterior probability. The first term measures the expected, i.e. weighted average of the Kulback-Leibler difference between the agent's beliefs about external states given the observed sensory data and the policy ($Q(\tilde{h}|\tilde{s},\pi)$) and the beliefs about external states only given the policy ($Q(hP|\pi)$) [...]. In contrast, the second term $EQ(\tilde{s}|\pi)\ [\log P(\tilde{s}|C)]$ represents the expected likelihood of observing sensory data ($\tilde{s}$) given a set of preferences (C).
>
> Minimizing this term encourages the agent to perform actions that will produce expected observations $\tilde{s}$. Accordingly, this term can be associated with desire-like states whose role is to bring about external states that would produce preferred observational outcomes. (Jurjako, 2026, p. 277)

An important part of what Jurjako says here seems to me to be absolutely correct, namely, the part that has to do with personal-level propositional attitudes. Although the equation's representational primitives, so to speak, are couched in decision-theoretic and information-theoretic terms, it is very plausible that these can be taken to operationalize personal-level doxastic states (in the first term) and personal-level conative states (in the second term).[27] Jurjako interprets the equation as formalizing what is often described as the exploration/ exploitation trade-off (between information gain, captured in the first term, and pragmatic cost, captured in the second). This is insightful.

The conclusion that he draws from this insight, however, is ambiguous. He concludes that the active inference framework has resources to capture cognitive states, processes, and abilities as they are construed in folk-psychological explanations. This is correct in the sense that the equation, and hence by extension the active inference approach in general, can incorporate folk psychological states in ways that are genuinely predictive and explanatory. But he also appears to take 'captures' in a much stronger sense, namely, as incorporating them in an algorithm that computes (solves) the problem of minimizing free energy. And this seems unjustified, because the simple fact of writing a term in an equation that can, and perhaps should, be interpreted in terms of personal-level attitudes, does not make those attitudes *algorithmically tractable*.

Algorithmic tractability requires effective procedures that will take us from inputs of a certain type to outputs of a certain type. Marr's great achievement in his analysis of the visual system was outlining what such effective procedures might look like for the task of transforming the inputs to the visual system (changes in light intensity, and so forth) into the representation of three-dimensional objects. The Marr-Hildreth edge detection algorithm is a great example, because it provides a simple and effective way of identifying edges. In effect, it answers a Yes-No question in an invariant way for highly circumscribed types of input —in which respect it contrasts strikingly with the equation considered above, which has, in effect, a parameter corresponding to the agent's beliefs and another parameter corresponding to the agent's desires. To assimilate these two very different types of equation not only obscures important conceptual distinctions, it also obscures the magnitude and originality of Marr's achievement in showing how complex perceptual and cognitive phenomena can be elucidated algorithmically.

Jurjako has not identified any effective procedures comparable to the Marr-Hildreth edge detection algorithm in the case of the active inference research program (not surprisingly, if my comments above about the different explanatory projects at stake are correct). Perhaps, despite the reasons for doubt alluded to in section 1 above, some such effective procedures will be identified in the future. I submit, though, that up to now they have not yet been identified, and for that reason I feel comfortable continuing to maintain that the Marrian tri-level approach, because it depends so critically upon the notion of algorithmic computation, cannot be applied to the mind as a whole —and, moreover, that there remain significant conceptual difficulties for thinking that it might be applicable to any non-modular cognitive abilities.

REFERENCES

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Bermúdez, J. L. (1995). Ecological perception and the notion of a nonconceptual point of view. In J. L. Bermúdez, A. J. Marcel, & N. Eilan (eds.), *The body and the self* (pp. 153-173). MIT Press.

Bermúdez, J. L. (1998). *The Paradox of Self-Consciousness.* MIT Press.

Bermúdez, J. L. (2000). Naturalized sense data. *Philosophy and Phenomenological Research* 61, 353-374.

[27] For further discussion of the relation between decision theory and folk psychology see Bermúdez (2009).

Bermúdez, J. L. (2002). The sources of self-consciousness. *Proceedings of the Aristotelian Society, 102*(1), 87-107.

Bermúdez, J. L. (2003a). The elusiveness thesis, immunity to error through misidentification, and privileged access. In B. Gertler (ed.), *Self-knowledge and privileged access* (pp. 213-233). Ashgate.

Bermúdez, J. L. (2003b). *Thinking without words*. Oxford University Press.

Bermúdez, J. L. (2003c). The domain of folk psychology. In A. O'Hear (ed.), *Minds and persons* (pp. 1-29). Cambridge University Press.

Bermúdez, J. L. (2005a). Evans and the sense of "I". In J. L. Bermúdez (ed.), *Thought, reference, and experience: Themes from the philosophy of Gareth Evans* (pp. 164-94). Oxford University Press.

Bermúdez, J. L. (2005b). *Philosophy of psychology: A contemporary introduction*. Routledge.

Bermúdez, J. L. (2005c). The phenomenology of bodily awareness. In D.W. Smith, A.L. Thomasson (eds.), *Phenomenology and philosophy of mind* (pp. 259-317). Oxford University Press.

Bermúdez, J. L. (2009). *Decision theory and rationality*. Oxford University Press.

Bermúdez, J. L. (2010). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press.

Bermúdez, J. L. (2011a). Bodily awareness and self-consciousness. In S. Gallagher (ed.), *Oxford handbook of the self* (pp. 157-179). Oxford University Press.

Bermúdez, J. L. (2011b). Self-knowledge and the sense of "I". In A. Hatzimoysis (Ed.). *Self-knowledge* (pp. 236-245). Oxford University Press.

Bermúdez, J. L. (2011c). The forcefield puzzle and mindreading in nonhuman primates. *Review of Philosophy and Psychology*, *2*(3), 397-410.

Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind (2nd edition)*. Cambridge University Press.

Bermúdez, J. L. (2015). Bodily ownership, bodily awareness and knowledge without observation. *Analysis, 75*(1), 37-45.

Bermúdez, J. L. (2017a). Ownership and the space of the body. In A. Alsmith & F. de Vignemont (eds.), *The body and the self revisited* (pp. 117-143). MIT Press.

Bermúdez, J. L. (2017b). *Understanding "I": Language and thought*. Oxford University Press.

Bermúdez, J. L. (2017c). Yes, essential indexicals really are essential*. *Analysis, 77*(4), 690-694.

Bermúdez, J. L. (2018a). Frames, rationality, and self-control. In J. L. Bermúdez (Ed.), *Self-control, decision theory, and rationality*. Cambridge University Press.

Bermúdez, J. L. (2018b). The bodily self, commonsense psychology, and the springs of action. In J. L. Bermúdez (Ed.), *The bodily self* (pp. 183-202). MIT Press.

Bermúdez, J. L. (2018c). Bodily ownership, bodily awareness, and knowledge without observation. In J. L. Bermúdez, *The bodily self* (pp. 183-202). MIT Press.

Bermúdez, J. L. (2018d). Inner speech, determinacy, and thinking consciously about thoughts. In P. Langland-Hassan (ed.), *Inner speech* (pp. 199-220). Oxford University Press.

Bermúdez, J. L. (2018e). *The bodily self: Selected essays on self-consciousness*. MIT Press.

Bermúdez, J. L. (2019a). First person thoughts: Shareability and symmetry. *Grazer Philosophische Studien, 96*(4), 629-638.

Bermúdez, J. L. (2019b). Framing as a mechanism for self-control. In A. Mele (ed.), *Surrounding self-control* (pp. 361-383). Oxford University Press.

Bermúdez, J. L. (2020). *Frame it again: New tools for rational thought*. Cambridge University Press.

Bermúdez, J. L. (2022a). Rational framing effects: A multidisciplinary case. *Behavioral and Brain Sciences*, 45, e220, 1-59

Bermúdez, J. L. (2022b). Frames and rationality: Response to commentators. *Behavioral and Brain Sciences*, 45, e248

Bermúdez, J. L. (2025). *Aristotle's* De Anima*: a guide*. Oxford University Press.

Bernués, D. & Jorba, M. (2026). Bermúdez's view on inner speech: A critical assessment. *Theoria. An International Journal for Theory, History and Foundations of Science, 41*(2), 215-230. https://doi.org/10.1387/theoria.26311

Brown, D. H. (2026). From immediate perception to perceptual belief. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 157-178. https://doi.org/10.1387/theoria.26036

Byrne, A. (2018). *Transparency and self-knowledge*. Oxford University Press.

Campbell, J. (1994). *Past, space, and self*. MIT Press.

Cappelen, H., & Dever, J. (2013). *The inessential indexical: on the philosophical insignificance of perspective and the first person*. Oxford University Press.

Carruthers, P. (2018). The causes and contents of inner speech. In P. Langland-Hassan (ed.), *Inner speech* (pp. 199-220). Oxford University Press.

Caston, V. (2002). Aristotle on consciousness. *Mind, 111*(444), 751-815.

Evans, G. (1982). *The varieties of reference*. Oxford University Press.

Dohrn, D. (2026). Can quasi-cyclical preferences be rational?. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 231-244. https://doi.org/10.1387/theoria.26006

Dummett, M. (1973). *Frege: philosophy of language*. Duckworth.

Dummett, M. (1981). *The interpretation of Frege's philosophy*. Duckworth.

Fernandez, J. (2013). *Transparent minds: a study of self-knowledge*. Oxford University Press.

Fisher, S. (2026). Discursive deadlock and the social (media) life of frames. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 245-261. https://doi.org/10.1387/theoria.26260

Frege, G. (1918/1977). 'Der Gedanke'. Translated in P. T. Geach (ed.), *Frege: Logical Investigations*. Basil Blackwell.

Friston, K., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., . . . Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports, 1024*, 1-29.

Friston, K., Sengupta, B., & Auletta, G. (2014). Cognitive dynamics: From attractors to active inference. *Proceedings of the IEEE, 102*(4), 427-445.

Gallagher, S. (2026). Why a perceptual zero-point amounts to nothing. *THEORIA. An International Journal for Theory, History and Foundations of Science*, 41(2), 179-195. https://doi.org/10.1387/theoria.26306

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton MifIlin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton MifIlin. Hume, D. (1739/1888), *A treatise of human nature* (ed. By L. A. Selby-Bigge). Clarendon Press.

Guala, F. (2022). Consistent preferences, conflicting reasons, and rational evaluations. *Behavioral and Brain Sciences 45*, 27-28.

Husserl, E. (1989) *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy; Second book: studies in the phenomenology of constitution*. (R. Rojcewicz & A. Schuwer, Trans.). Kluwer.

Husserl, E. (2006). *The basic problems of phenomenology: from the Lectures, Winter Semester, 1910-1911.* (I. Farin & J. G. Hart, Trans.). Springer.

Jurjako, M. (2026). How general are Marr's levels of analysis?: An assessment of Bermúdez's view. *THEORIA. An International Journal for Theory, History and Foundations of Science*, 41(2), 262-282. https://doi.org/10.1387/theoria.25896

Kahneman, D., & Tversky, A. (2000). *Choices, values, frames*. Cambridge University Press.

Lakoff, G. (2004). *Don't think of an elephant: Know your values and frame the debate*. Chelsea-Green Publishing.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Martínez-Manrique, F., & Vicente, A. (2010). What the...! The role of inner speech in conscious thought. *Journal of Consciousness Studies*, *17*(9-10), 141-167.

Peacocke, C. (1999). *Being known*. Oxford University Press.

Perry, J. (1979). The problem of the essential indexical. *Noûs, 13*, 3-21.

Pettigrew, R. (2022). Competing reasons, incomplete preferences, and framing effects. *Behavioral and Brain Sciences* 45, 35-6.

Ross, W. D. (1984). Nicomachean Ethics. In J. Barnes (ed.), *The complete works: the Rev. Oxford translation* (pp. 1729-1867). Princeton University Press.

Sainsbury, R. M. (1998). Indexicals and reported speech. *Proceedings of the British Academy, 95*, 49-69.

Serrahima, C. (2026). Self-consciousness, ψ and ϕ. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 196-214. https://doi.org/10.1387/theoria.26190

Shields, C. (2016). *Aristotle's De Anima: translated with an introduction and commentary*. Clarendon Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, 59(4), S251-S278.

Valente, M. (2026). Asymmetries between 'you' and 'I'. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 136-156. https://doi.org/10.1387/ theoria.26508

Verdejo, V. M. (2026). Thoughts about oneself to share in context: Meeting Bermúdez's challenge. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *41*(2), 122-135. https://doi.org/10.1387/theoria.26297

Wittgenstein, L. (1921/1961). *Tractatus Logico-Philosophicus*, D. F. Pears and B. F. McGuinness, Trans.), Humanities Press.

Wittgenstein, L. (1958). *Preliminary studies for the "Philosophical investigations," generally known as the Blue and Brown Books*. Basil Blackwell.

José Luis Bermúdez is Professor of Philosophy and Charles H. Gregory '64 Chair in Arts and Sciences at Texas A&M University.

Address: Suite 409, Academic Building, Texas A&M University 77843 and Department of Philosophy, Texas A&M University, College Station TX 77843-4223. Email jbermudez@tamu.edu –ORCID: 0000-0002-8312-1013