

THEORIA

Vol. 41/2 • May 2026

ESTABLISHED IN 1952 BY MIGUEL SÁNCHEZ-MAZAS
Second Series

An International Journal for Theory, History and Foundations of Science



Ὁ Θεὸς ἀριθμεῖται

CALIJ

Centro de Análisis, Lógica e Informática Jurídica (CALIJ)



<https://ojs.ehu.eus/index.php/THEORIA>

THEORIA

REVISTA DE TEORÍA, HISTORIA Y FUNDAMENTOS DE LA CIENCIA

AN INTERNATIONAL JOURNAL FOR THEORY, HISTORY AND FOUNDATIONS OF SCIENCE

ESTABLISHED in 1952 by MIGUEL SÁNCHEZ-MAZAS Second Series

EDITORIAL BOARD

Editor-in-chief: Andoni IBARRA (University of the Basque Country, EHU)

Editors: Michele PALMIRA (Universidad Complutense de Madrid), Javier GONZÁLEZ DE PRADO (UNED)

Logic and philosophy of logic and mathematics: Eduardo A. BARRIO (Universidad de Buenos Aires)

Philosophy of language: Delia BELLERI (University of Lisbon)

Philosophy of mind and cognition: Marta JORBA (Universitat Pompeu Fabra)

Epistemology: Michele PALMIRA (Universidad Complutense de Madrid)

Philosophy of science: José DíEZ (Universitat de Barcelona)

History of science & Science and Technology Studies: Miguel GARCÍA-SANCHO (University of Edinburgh)

Consulting Editor: Cristina CORREDOR (Universidad de Valladolid)

Managing Editor: Mario SANTOS-SOUSA (University College London)

Editorial Assistant: Jorge GUARDIOLA BLEDA (Universitat de València)

ADVISORY BOARD

Juan José ACERO (†) (Universidad de Granada), Ignazio ANGELELLI (†) (University of Texas at Austin), Manuel ATIENZA (Universitat d'Alacant), Salvador BARBERÀ (Universitat Autònoma de Barcelona), Nancy CARTWRIGHT (Durham University), Newton C.A. da COSTA (†) (Universidade de São Paulo), Joseph DAUBEN (City University of New York), Elías DÍAZ (Universidad Autónoma de Madrid), Javier ECHEVERRÍA (Ikerbasque & EHU), Josep Maria FONT (Universitat de Barcelona), Manuel GARCÍA-CARPINTERO (Universitat de Barcelona), Philip KITCHER (Columbia University), Bruno LATOUR (†) (Sciences Po, Paris), Larry LAUDAN (†) (UNAM), Javier DE LORENZO (Universidad de Valladolid), Thomas MORMANN (EHU), C. Ulises MOULINES (Ludwig-Maxilians-Universität München), Carlos MOYA (Universitat de València), Javier MUGUERZA (†) (UNED), Roberto TORRETTI (†) (Universidad de Puerto Rico), Enric TRILLAS (Universidad Politécnica de Madrid), Nicanor URSUA (EHU), Bas C. VAN FRAASSEN (San Francisco State University)

EDITORIAL OFFICE

Editorial manager: Ainhoa FERNÁNDEZ (CALIJ, Donostia/San Sebastián, Spain)

Mailing address: CALIJ-THEORIA, Elhuyar plaza 2 (20018 Donostia/San Sebastián, Spain)

Tel.: (+34) 943 017 447. Fax: (+34) 943 015 990. E-mail: editor.theoria@ehu.eus

THEORIA is a four-monthly journal (issues in January, May and September). Accepted articles are published in Advanced Access as soon as they are ready, prior to their assignment to a specific issue. The contents of this journal are covered, among others, by the following abstracting and indexing services: Arts & Humanities Citation Index®, ISI Alerting Services, Current Contents® / Arts & Humanities, Bulletin Signalétique 519, DICE, ICYT, ISOC and RESH of the Spanish research center CSIC, Mathematical Reviews, Current Mathematical Publications, MathSci, Philosopher's Index, Repertoire bibliographique de la Philosophie, and SCOPUS.

THEORIA's website

<http://www.ehu.eus/theoria>

CONTENTS

Volume 41/2, May 2026, pp. 111-341

ISSN 0495-4548

e-ISSN 2171-679X

Monographic Issue

XXX SIUCC – José Luis Bermúdez

Guest editors: Matheus Valente, Víctor M. Verdejo and Valentine Simpson

Valentine SIMPSON, Matheus VALENTE

The philosophy of José Luis Bermúdez: an introduction / *La filosofía de José Luis Bermúdez: una introducción* . . . 113

Víctor M. VERDEJO

Thoughts about oneself to share in context: meeting Bermúdez's challenge / *Pensamientos sobre uno mismo para compartir en contexto: una respuesta al reto de Bermúdez*. 122

Matheus VALENTE

Asymmetries between 'you' and 'I' / *Asimetrías entre 'tú' y 'yo'* 136

Derek H. BROWN

From immediate perception to perceptual belief / *De la percepción inmediata a la creencia perceptiva* 157

Shaun GALLAGHER

Why a perceptual zero-point amounts to nothing / *Por qué un punto cero perceptivo no significa nada*. 179

Carlota SERRAHIMA

Self-consciousness, ψ and ϕ / *Autoconsciencia, ψ y ϕ* 196

Daphne BERNUÉS MOSS, Marta JORBA GRAU

Bermúdez's view on inner speech: a critical assessment / *La teoría del habla interna de Bermúdez: una evaluación crítica* 215

Daniel DOHRN

Can quasi-cyclical preferences be rational? / *¿Pueden ser racionales preferencias cuasi-cíclicas?* 231

Sarah A. FISHER

Discursive deadlock and the social (media) life of frames / *Bloqueo discursivo y la vida social (mediática) de los marcos* . . . 245

Marko JURJAKO

How general are Marr's levels of analysis? An assessment of Bermúdez's view / *¿Cómo de generales son los niveles de análisis de Marr? Una evaluación de las propuestas de Bermúdez* 262

José Luis BERMÚDEZ

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*. 283

SUMMARY 340

Monographic Issue

XXX SIUCC – JOSÉ LUIS BERMÚDEZ

GUEST EDITORS: MATHEUS VALENTE,
VÍCTOR M. VERDEJO AND VALENTINE SIMPSON



THE PHILOSOPHY OF JOSÉ LUIS BERMÚDEZ: AN INTRODUCTION

(*La filosofía de José Luis Bermúdez: una introducción*)

Valentine Simpson*

Universitat de València

<https://orcid.org/0009-0003-2263-8857>

Matheus Valente**

UNED

<https://orcid.org/0000-0001-6380-2623>

Few philosophers working today have ranged as widely and as productively across the intersecting territories of philosophy of mind, philosophy of language, philosophy of psychology and cognitive science, epistemology and the theory of rationality as José Luis Bermúdez. Over a career spanning more than three decades, Bermúdez has offered (and continues to offer) sustained, original, and technically rigorous contributions to questions about —to cite just a few examples— the nature of self-consciousness, the relationship between perception and cognition, mental representation and nonconceptual content, the first-person in both language and thought, decision theory and the rationality of preferences, bodily awareness (the list goes on).¹

This special issue of *THEORIA* brings together nine original contributions that engage critically with various strands of José Luis Bermúdez's work, along with responses by Bermúdez himself. The contributions are grouped into six thematic clusters, and are followed by Bermúdez's responses²: (1) **First-Person Thought** (Víctor M. Verdejo; Matheus Valente), (2) **Sense Data** (Derek H. Brown), (3) **Bodily Awareness and Self-Consciousness** (Shaun Gallagher; Carlota Serrahima), (4) **Inner Speech** (Daphne Bernués & Marta Jorba), (5) **Frames, Preferences, and Rationality** (Daniel Dohrn; Sarah A. Fisher), (6) **Explanation in Cognitive Science** (Marko Jurjako). What follows is a brief description

¹ In 2025 alone, Bermúdez released two new books – the Oxford guide to Aristotle's *De Anima* (2025c) and the edited collection *Sharing thoughts: philosophical perspectives on intersubjectivity and communication* (2025a; co-edited with Víctor M. Verdejo and Matheus Valente) – whereas a third one – *The self in premodern thought: from antiquity to the renaissance in europe* (co-edited with Catherine Conybeare) – is forthcoming with Cambridge University Press.

² Bermúdez's responses are grouped together in a single piece (Bermúdez 2026), structured in six sections, each corresponding to the thematic clusters mentioned above.

* **Correspondence to:** Valentine Simpson. Valencia Philosophy Lab, Departament de Filosofia, Facultat de Filosofia i Ciències de l'Educació, Universitat de València, Av. Blasco Ibáñez 30, Valencia 46010, Spain – valensimpson@gmail.com – <https://orcid.org/0009-0003-2263-8857>

** **Correspondence to:** Matheus Valente. UNED. Humanidades. Paseo de Senda del rey, 7 (28040 Madrid, Spain) – matheusvalenteleite@gmail.com – <https://orcid.org/0000-0001-6380-2623>

How to cite: Simpson, Valentine; Valente, Matheus (2026). «The philosophy of José Luis Bermúdez: an introduction»; *Theoria. An International Journal for Theory, History and Foundations of Science*, 41(2), 113-121. (<https://doi.org/10.1387/theoria.28722>).

Received: 22 April, 2026; Final version: 19 June, 2026.

ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

of each cluster and Bermúdez's responses. Far from aiming at exhaustiveness, these summaries are intended simply as an invitation to delve into the exchanges between the contributors and Bermúdez that unfold in the papers and responses collected in this issue.

1. First-Person Thought

First-person thought, a recurring thread in Bermúdez's work since its early beginnings, is given its most systematic presentation by that author in the monograph *Understanding "I": language and thought* (2017a), the primary target of the first two papers collected in this issue by Víctor M. Verdejo and Matheus Valente.³ In particular, both Verdejo and Valente address Bermúdez's defense of the shareability of first-person thoughts (i.e. the claim that, in ordinary conversational contexts, the same thoughts expressible by a subject with 'I' can be expressed by others with 'you'). Indeed, one of the centre pieces of Bermúdez's monograph is a battery of arguments aiming to show that, contrary to what Frege (1918) himself once suggested (as well as more recent neo-Fregean authors such as Evans (1981) and Morgan (2009))⁴, the shareability of first-person thoughts is perfectly compatible with a proper account of their peculiar cognitive and epistemic features.

In *Thoughts about oneself to share in context: meeting Bermúdez's challenge* (Verdejo 2026), Víctor M. Verdejo presents the latest instalment in a critical exchange with Bermúdez which started with Verdejo (2018) and was followed by Bermúdez (2019). Overall, Verdejo is highly sympathetic to the spirit of Bermúdez's project, duly acknowledging the importance of vindicating the shareability of first-person thoughts, though they do disagree about how this ought to be cashed out, as well as the consequences of this view for the individuation of thought. Whereas Bermúdez explains what it is for uses of 'I' and 'you' to express the same (token) thought in terms of the relevant subjects' deploying the same practical capacity for uniquely locating an individual in space-time, Verdejo argues that the same goal can be accomplished in terms of the relevant subjects' deploying thoughts that are constituted by the same reference-rule. To be sure, Verdejo agrees with Bermúdez that the relevant reference-rule cannot be of a traditional token-reflexive kind (e.g. a token-thought involving 'I' refers to its thinker), for then it would not have been possible for the same token-thoughts to be expressible by uses of 'you'. With this in mind, Verdejo proposes an alternative conception of the reference-rule of first-person thoughts, **the self rule**, which determines the referent of these thoughts as *the subject* of the relevant thought. Unlike token-reflexive rules which always refer back to the thinker, the self rule, Verdejo tells us, is perfectly compatible with the shareability of first-person thoughts, as not only the thinker of the relevant thought may be aware of herself as the subject of said thought, but also her interlocutors. This invites us to carefully reconsider the role of the so-called Fregean Criterion in the individuation of first-person thought. But one might take this proposal to ultimately depend on our being able to give a substantial account of what it is for one to be *the subject* of a thought over and above being the thought's referent or the thought's producer, a difficult task the prospects of which Bermúdez evaluates in his response: *Symmetry, privacy, and "I"-thoughts: A reply to Valente and Verdejo*.

Valente's (2026) contribution —*Asymmetries between 'you' and 'I'*— is more pessimistic than Verdejo's regarding the prospects of vindicating first-person thoughts' shareability within a broadly Fregean framework. The central part of this paper revolves around an argument supposed to entail that, if first-person thoughts are shareable, then they cannot play their fundamental role in explanations of rational action and production of motivating reasons. This is a serious accusation given that, not only Bermúdez, but anybody who sympathises with the need to distinguish between sense and reference will be prone to accept a constitutive connection between thought and rational action.⁵ In rough terms, the key

³ See also Bermúdez (2005b, 2011b, 2017b, 2019).

⁴ Frege once suggested that first-person thoughts are incommunicable given that they are based on a primitive way in which each subject is presented to oneself but to no one else. Both Evans (1981) and Morgan (2009) develop this type of reasoning and ultimately subscribe to the privacy of first-person thoughts.

⁵ Valente's argument harks back to Perry's (1977) notorious bear attack scenario.

idea is that two subjects might have distinct reasons for action even if they share all of each other's third-person beliefs and desires. If first-person thoughts were shareable, Valente argues, then the difference between these subjects' reasons would be left without any plausible attitudinal explanation. Valente's argument crucially relies on the assumption that motivating reasons supervene on nothing but agents' attitudinal states. However, one may fairly wonder whether this restrictive picture leaves out anything of importance. In his response, Bermúdez emphasises the role of agents' occurrent states (e.g. perceptions) for the production of rational action and, noticing this relevant issue's omission from Valente's discussion, deploys it in an attempt to defuse his argument.

2. *Sense Data*

Bermúdez's work in the philosophy of perception stands tall among his philosophical achievements. This most obviously includes a set of highly influential papers on the notion of nonconceptual content. Slightly less well-known is his examination of the prospects for a naturalistic treatment of sense data in his 2000 paper *Naturalized sense data*, the main target of the third contribution to this issue, Derek H. Brown's *From immediate perception to perceptual belief* (Brown 2026).⁶

Bermúdez's (2000) *naturalized sense datum theory* (NSD) differs from traditional *sense datum* theory (SDT) in taking the immediate objects of perception, i.e. *sense data*, to be physical objects – though not the familiar three-dimensional objects to which we routinely refer in demonstrative thought and speech, but their facing surfaces. Bermúdez's hypothesis was that the NSD is a sort of best of both worlds: it preserves the core insights that have historically lent plausibility to the SDT while avoiding the pitfalls of uncritical direct realism.

Brown is not convinced of Bermúdez's argument for the superiority of NSD over SDT in better accounting for the explanatory connection between our immediate objects of perception and our capacity for demonstrative reference to ordinary physical objects. He thus sets out both to show how, contrary to Bermúdez's claim, SDT is consistent with our capacity for acquiring (mediate) knowledge about the external world and to raise independent issues with NSD regarding its prospects for dealing with the Problem of Perception (i.e. explaining both the differences and the similarities between veridical, illusory, and hallucinatory perceptual experiences). Regardless of that, Brown closes his piece with a rejoinder friendly to Bermúdez, in noticing that many of NSD's resources will still be foundational to any theory – *sense datum*-based or not – of perception and of perceptual demonstratives.

In his response, Bermúdez attempts to tackle both horns of Brown's criticism, renewing his challenge to the defender of SDT's capacity to deploy the concept of deferred ostension in explaining demonstrative reference to ordinary physical objects as well as sketching a strategy with which the defender of NSD might tackle the Problem of Perception.

3. *Bodily Awareness and Self-Consciousness*

The bodily self: selected essays (Bermúdez, 2018a) remains an important landmark in the philosophical study of bodily awareness and its relation to self-consciousness, providing a comprehensive examination of the ways in which embodiment is constitutively connected to some of the most primitive forms of self-awareness in conscious beings.⁷ Though

⁶ It might be worth noticing that interest in Bermúdez's paper has been growing steadily, especially in recent years, most likely due to renewed interest in *sense datum* theories in contemporary philosophy. If Philpapers is right, it took close to ten years for Bermúdez's 2000 paper to receive its first citation. Notwithstanding that, the paper currently holds approximately thirty citations, almost all in publications from the last decade or so. Derek H. Brown's contribution to this issue along with Bermúdez's response are thus timely additions to this recent trend.

⁷ See also Bermúdez (1998, 2020b).

they approach this topic from different angles, this issue's contributions by Shaun Gallagher (2026) and by Carlota Serrahima (2026) fit together as explorations of Bermúdez's work on the body's involvement in self-consciousness.

Shaun Gallagher's contribution —*Why a perceptual zero-point amounts to nothing*— offers a probing critique of the commonly held view of the body as a sort of perceptual zero-point from which the world is egocentrically experienced and acted upon. His is a historically-informed, critical analysis of authors ranging from the phenomenological tradition of Husserl and Merleau-Ponty, to analytic philosophers of mind like Gareth Evans and Christopher Peacocke. Against them, Gallagher argues that the notion of a perceptual zero-point fails to carry the phenomenological weight assigned to it, and instead proposes a more complex view according to which the processes involved in perception and action must be mapped onto (different parts of) the body along varying axes and dimensions. Taking recourse from Bermúdez's own distinction between three different frames of bodily reference, as well as resources from the enactivist tradition, Gallagher sets forth a "thick" view of bodily experience based on agentive situations, which bring together agents *and* their environments. These situations are defined by actions that are structured by the subject's affordances, taking place in different time scales with varying degrees of intentionality.

Carlota Serrahima's paper —*Self-Consciousness, ψ and ϕ* — examines the distinctions between the psychological (ψ) and bodily (ϕ) dimensions of self-consciousness with the underlying aim of assessing the prospects for unified accounts of both phenomena, such as Bermúdez's own.⁸ One important insight of Serrahima's piece is that for both forms of ownership, ψ and ϕ , an equilibrium must be reached between two oppositely pulling ideas. Regarding ϕ -ownership, on the one hand, it seems there is an intimate link between experiences of bodily sensations and self-attributions of bodily ownership. On the other hand, empirical research points towards a break between them, as indicated by subjects reporting a lack of bodily ownership where bodily sensations are nonetheless present. ψ -ownership presents a parallel situation; authors maintain there is a compelling link between the presence of phenomenal states and a self which attributes said states. Similarly, however, empirical studies on phenomena such as cases of thought insertion point to a parallel break between occurrences of experiences and their self-attribution by subjects. In exploring the logical space for views on the senses of ownership respectively associated to ψ and ϕ self-consciousness, Serrahima argues that the positing of a distinctive phenomenology of ownership represents the best prospects for an integrated account of ψ and ϕ self-consciousness. As Serrahima argues, this raises a serious difficulty to Bermúdez's ambitions of giving a unified account of these phenomena while simultaneously endorsing the elusiveness of the self (while obviously *not* endorsing the analogous thesis of the elusiveness of the body).

Bermúdez's reply —*Elusiveness and the phenomenology of the self: Response to Gallagher and Serrahima*— engages with both the phenomenological and conceptual dimensions of these authors' challenges. In his response to Gallagher, Bermúdez sets out to clarify his own view on what is the central question regarding the notion of the body as a perceptual zero-point —i.e. not whether prereflective experience of the body is *of* a zero-point, as he takes Gallagher to presuppose, but instead: what it is like for experience to be *from* a perceptual zero-point— as well as to draw out some important interconnections between this issue and the elusiveness of the self. This allows Bermúdez to smoothly transition to his discussion of Serrahima's paper, wherein Humean worries about the self are discussed alongside Wittgenstein's distinction between the self-as-object and the self-as-subject, ultimately leading him to concede to Serrahima that his old account of bodily self-ascription ought to be revised in order to reconcile the elusiveness of the self with (much less elusive) bodily phenomenology.

4. Inner Speech

Daphne Bernués and Marta Jorba's (2026) joint contribution —*Bermúdez on inner speech: A critical assessment*— turns attention to Bermúdez's paper 'Inner speech, determinacy, and thinking consciously about thoughts' (2018b) where

⁸ Serrahima mostly focuses on Bermúdez's paper "The elusiveness thesis, immunity to error through misidentification, and privileged access", reprinted as chapter 4 in Bermúdez (2018a).

some of the key ideas developed in his monograph *Thinking without words* (2003) are applied to the phenomenon of inner speech. One of the central claims made by Bermúdez is that intentional ascent, the capacity to think about thoughts, requires semantic ascent, the capacity to think about words or linguistically represented thoughts. According to Bermúdez, inner speech is a key tool for semantic ascent. Therefore, inner speech is taken to be the main vehicle through which second-order thinking about thoughts takes place.

Bernués and Jorba subject Bermúdez's proposal to sustained scrutiny along two clearly distinct lines. First, they argue that his dependency thesis does not hold up in light of empirical research (citing, for example, studies indicating that children age 4-6, when inner speech is yet not developed, are capable of reflexive forms of thinking regardless). Second, they argue that it runs into an explanatory dilemma stemming from Bermúdez's view that a thought's canonical structure must be revealed in thinking about it through inner speech. Confronted with the semantic indeterminacy of natural language sentences,⁹ Bermúdez must hold that subpersonal computations deliver a determinate content apprehensible in inner speech. As Bernués & Jorba argue, this move leads to the following dilemma regarding inner speech's explanatory status: on the one hand, if content is determined subconsciously, it is not clear why we need expanded forms of inner speech and not highly condensed ones (which presumably would not reveal its canonical structure), rendering expanded inner speech superfluous; on the other hand, if we instead accept the need for these expanded forms of inner speech, the problem becomes that of determining the form of the proposition being tokened. We may only be able to arrive at the right proposition by a revelation of its logical form through formal means, but that is clearly not what is entertained in inner speech, in which case inner speech is inaccurately described. Bernués & Jorba's contribution ends with a proposal amenable to Bermúdez's insights: while inner speech may be *one* way of helping make thoughts conscious, there may be other ways, such as certain forms of unsymbolized thinking.

Bermúdez responds in *Inner speech and intentional ascent: Reply to Bernués and Jorba*. In this piece, he starts by suggesting there be alternative interpretations of the empirical data put forth in their paper, alluding to a distinction between intentional ascent properly speaking and more primitive analogues of that capacity. However, the bulk of Bermúdez's response consists in his attempt to tackle both horns of Bernués & Jorba's explanatory dilemma, one whose force and originality Bermúdez is quick to acknowledge, a task which ultimately requires him to submit aspects of his previous view to revision.

5. *Frames, Preferences, and Rationality*

Bermúdez's latest monograph, *Frame it again: new tools for rational choice* (2020a), sees that author engaging with fundamental disputes in decision theory and, more broadly, the theory of rationality.¹⁰ Among its important contributions is the development of a sophisticated account of practical rationality which takes framing phenomena and framing effects as integral and ineliminable elements of that type of reasoning—as opposed to lesser-important deviations of ideal rationality, as the orthodoxy in decision theory has traditionally maintained. One upshot of Bermúdez's proposal is that, due to framing-sensitivity, perfectly rational agents might have quasi-cyclical preferences whereby they prefer A to B, and B to C, while knowing full well that A and C are the same option differently framed.

Daniel Dohrn's paper in this issue —*Can quasi-cyclical preferences be rational?*— challenges some of Bermúdez's most fundamental claims: whether preferences are frame-relative, whether there are in fact such quasi-cyclical preferences, and whether these could play the role in rational deliberation that Bermúdez claims they do. It seems, Dohrn tells us, that a plausible necessary condition for preference is that one consciously, deliberately, and voluntarily acts on one only

⁹ The objection from semantic indeterminacy is due to Martínez-Manrique & Vicente (2010), a work which plays an important role in both Bernués & Jorba's contribution and Bermúdez's response.

¹⁰ See also Bermúdez (2011a, 2022a, 2022b).

if one does not have an overriding preference. Quasi-cyclical preferences do not satisfy this condition, as one will not be able to act on a quasi-cyclical preference consciously, deliberately, and voluntarily given the existence of an overriding preference existing under a different frame. We also find a problem of attribution of such quasi-cyclical preferences: it is not clear why we need to posit them to explain a subject's overt actions instead of using more orthodox tools, such as wishes or desires. More importantly, perhaps, is the role for rationality that Bermúdez wants to attribute to these preferences. Dohrn makes an illuminating distinction between two ways a preference may be labelled as rational: that of *input*-rationality, concerning the processes that influence decision-making, and that of *output*-rationality, which properly concerns the action-guiding role of a preference. We may accept that quasi-cyclical preferences may hold a rational role on the input side, as the practice of a subject's framing and reframing their preferences may be a good way of responding to some complex situation. However, dove-tailing his answer to the question concerning their status, Dohrn concludes that quasi-cyclical preferences cannot count as rational on the output side given the tension between one preference's (under a frame) guiding a subject's action, and there being another preference (under a different frame) that overrides the former. Dohrn ends the piece by demonstrating this conclusion as applied to three applications of Bermúdez's framework: time-inconsistencies, game theory, and interpersonal conflicts.

As a nice continuation of the previous contribution, Sarah Fisher's (2026) paper —*Discursive deadlock and the social (media) life of frames*— engages directly with the question of how the framing framework can both account for, and give solutions to situations of interpersonal conflict or, as Bermúdez also calls them, of discursive deadlock. In short, Bermúdez's claim is that they are situations in which subjects' frames clash with each other. As a solution, subjects ought to engage in frame-sensitive reasoning, ideally adopting one another's frames to better understand each other. Fisher pushes an objection against the very foundation of Bermúdez's proposal: that frames do not (always) put subjects in touch with *reasons*, but instead may also serve as signals of identity and social affiliation. She draws an interesting comparison with the phenomena of slurs and dogwhistles, as two cases where linguistic devices also function for the purpose of signaling identity and affiliation. The upshot of this alternative function of frames is two-fold. First, that discursive deadlock may be caused less by the opposing reasons that subjects entertain under their respective frames, and more by the social identity that they intend to present via said frames. Second, that Bermúdez's prescription for the adoption of one another's frames may be neither practical nor effective. Fisher finally turns to social media to highlight this point, as a case study of the way users may be compelled, not to perform the frame-relative due diligence Bermúdez calls for, but to remain in their identity-building and identity-sustaining frames. In this context, discursive deadlock, it is concluded, may be more readily dealt with through changes in the very structure of (online) social interactions.

Bermúdez responds to both papers in *Preferences, frames, and reasons: Reply to Dohrn and Fisher*. His response to Dohrn focuses primarily on defending his original position. Against Dohrn's scepticism, Bermúdez maintains that quasi-cyclical preferences do not reflect irrationality but arise from the way objects of preference are presented, altogether defending the possibility of rational framing effects more generally. Turning to Fisher, Bermúdez addresses her concerns about the normative ambitions of frame-sensitive reasoning, particularly its role in resolving discursive deadlock. While acknowledging Fisher's point that frames can function as signals of identity and affiliation, he argues that this does not undermine their potential to structure reasoning and facilitate progress in difficult deliberative contexts.

6. Explanation in Cognitive Science

In *Philosophy of psychology: a contemporary introduction* (2005a) and the first two editions of the textbook *Cognitive science: an introduction to the science of the mind* (2010, 2014), Bermúdez argued for significant limits on the applicability of Marr's (1982) tri-level analysis which analyzes cognitive systems at computational, algorithmic, and implementational levels. Specifically, Bermúdez maintained that this framework is probably only applicable to modular systems but not to the mind as a whole.

Marko Jurjako's (2026) contribution to this issue —*How general are Marr's levels of analysis?*— reexamines these claims in the light of recent theoretical developments in active inference and the free energy principle. The piece follows a clear argumentative strategy: Jurjako first provides reasons to think that, *contra* Bermúdez, Marr's tri-level analysis can in fact be applied at the non-modular level; then, through an appeal to the active inference model, which Jurjako shows how to map onto the Marrian levels of analysis, he presents a positive reevaluation of the applicability of the tri-level analysis to the mind as a whole. In rough terms, the active inference framework rests on the free energy principle, which claims that self-organising systems work to reduce the amount of environmental surprise in order to maintain their physical integrity. The general purpose of an intelligent mind, as Jurjako argues, may be seen as that of minimising uncertainty (and minimising loss of energy). Importantly, the application of the free energy principle to the study of the mind promises to provide a bridge between personal-level and subpersonal-level explanation (or even to blur their distinction).

Bermúdez responds to Jurjako in *Active inference, the Free Energy Principle, and Marr's tri-level hypothesis: Reply to Jurjako*. While Bermúdez grants that the free energy principle may indeed help illuminate the personal/subpersonal interface, he is more sceptical about its fitting the requirements of the essentially top-down Marrian approach, such as the identification of algorithmically tractable effective procedures linking inputs to outputs. For this reason, Bermúdez maintains his opinion that Marr's tri-level analysis cannot be applied to the mind as a whole, though he acknowledges the possibility that further developments in these areas might lead him to revise his conviction.

7. Concluding Remarks

We hope readers will indulge us in repeating our opening statement; we believe that, together, the diverse and compelling contributions gathered in this special issue serve as a powerful testament to both the extensive reach and the profound intellectual fertility of José Luis Bermúdez's philosophical vision. The papers collected here span a considerable spectrum, ranging from friendly extensions of his theoretical frameworks to incisive critical challenges that push the boundaries of his views. Yet Bermúdez's responses display an equal range of tact, from firm defenses of his original views, to acknowledgments of their limitations and some proposed revisions. Along with its dialectical nature, the special issue's sheer breadth of topics marks out a remarkably wide and rich terrain for contemporary analytic philosophy.

We wish to express our profound gratitude to all the contributors for their exceptionally stimulating and thought-provoking papers, which have enriched this volume immensely. Our thanks are equally extended to Professor Bermúdez for his generous expenditure of time and intellectual effort in providing thorough and insightful engagement with each and every contribution. We also wish to convey our sincere appreciation to the editors of *THEORIA* for their support throughout the entire duration of this project. Furthermore, we commend *THEORIA*'s continued commitment to ensuring that cutting-edge philosophical work is made freely and openly accessible to the widest possible global audience.

Acknowledgements

This output has received financial support via the grant RYC2021-033972-I from MCIU/AEI/10.13039/501100011033 and the European Union "NextGenerationEU"/PRTR; the grant PID2023-151071NB-I00 from MCIU/AEI/10.13039/501100011033/FEDER, UE; the grant "Communication, Inner Speech and MentalHealth" (Ref. CIH-PD4A7507) from the Fundación Ramón Areces; the project "Semántica y epistemología del pensamiento *de se*" (Investigadores Jóvenes UNED 2026); and the project "Conceptual Engineering and Communicative Disruptions: between Semantics and Pragmatics" (grant PID2024-157224NB-I00).

REFERENCES

- Bermúdez, J. L. (1998). *The paradox of self-consciousness: Representation and mind*. MIT Press.
- Bermúdez, J. L. (2000). Naturalized sense data. *Philosophy and Phenomenological Research*, 61(2), 353-374.
- Bermúdez, J. L. (2003). *Thinking without words*. Oxford University Press.
- Bermúdez, J. L. (2005a). *Philosophy of psychology: a contemporary introduction*. Routledge.
- Bermúdez, J. L. (2005b). Evans and the sense of 'I'. In J. L. Bermúdez (Ed.), *Thought, Reference, and Experience: Themes from the Philosophy of Gareth Evans* (pp. 164-194). Oxford University Press.
- Bermúdez, J. L. (2010). *Cognitive science: an introduction to the science of the mind* (1st ed.). Cambridge University Press.
- Bermúdez, J. L. (2011a). *Decision theory and rationality*. Oxford University Press.
- Bermúdez, J. L. (2011b). Self-knowledge and the sense of 'I'. In A. Hatzimoysis (Ed.), *Self-Knowledge* (pp. 226-245). Oxford University Press.
- Bermúdez, J. L. (2014). *Cognitive science: an introduction to the science of the mind* (2nd ed.). Cambridge University Press.
- Bermúdez, J. L. (2017a). *Understanding "I": Language and thought*. Oxford University Press.
- Bermúdez, J. L. (2017b). Yes, essential indexicals really are essential. *Analysis*, 77(4), 690-694.
- Bermúdez, J. L. (2018a). *The bodily self: Selected essays*. MIT Press.
- Bermúdez, J. L. (2018b). Inner speech, determinacy, and thinking consciously about thoughts. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices* (pp. 199-220). Oxford University Press.
- Bermúdez, J. L. (2019). First person thoughts: shareability and symmetry. *Grazer Philosophische Studien*, 96(4), 629-638.
- Bermúdez, J. L. (2020a). *Frame It Again: New Tools for Rational Choice*. Cambridge University Press.
- Bermúdez, J. L. (2020b). Four theses about self-consciousness and bodily experience: Descartes, Kant, Locke, and Merleau-Ponty. *Journal of the American Philosophical Association*, 6(1), 96-116.
- 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., Valente, M. & Verdejo, V. M. (eds.) (2025a). *Sharing thoughts: Philosophical perspectives on intersubjectivity and communication*. Oxford University Press.
- Bermúdez, J. L. (2025b). Frames, senses, and thought-equivalence. In J. L. Bermúdez, M. Valente & V. M. Verdejo (Eds.), *Sharing thoughts: Philosophical perspectives on intersubjectivity and communication* (pp. 75-91). Oxford University Press.
- Bermúdez, J. L. (2025c). Aristotle's *De Anima*. Oxford University Press.
- Bermúdez, J. L. (2026). Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Bernués and Jorba, Gallagher, Serrahima, Brown, 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>
- Bermúdez, J. L., & Conybeare, C. (forthcoming) *The self in premodern thought: From antiquity to the renaissance in Europe*. Cambridge 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>
- 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>
- Evans, G. (1981). Understanding demonstratives. In H. Parret & J. Bouveresse (Eds.), *Meaning and Understanding* (pp. 280-303). De Gruyter.
- Frege, G. (1918/1977). Der Gedanke. Translated in P. T. Geach (Ed.), *Frege: Logical Investigations* (pp. 1-30). Basil Blackwell.
- Fisher, S. A. (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>
- 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>
- 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>
- Martínez-Manrique, F. & Vicente, A. (2010). What the...! The role of inner speech in conscious thought. *Journal of Consciousness Studies* 17 (9-10), 141-67.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Morgan, D. (2009). Can you think my 'I'-thoughts?. *Philosophical Quarterly* 59(234), 68-85.
- Perry, J. (1977). Frege on demonstratives. *Philosophical Review* 86(4), 474-497.
- 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>
- 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. (2018). On having the same first person thought. *Grazer Philosophische Studien*, 95(4), 566-587.
- 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>

VALENTINE SIMPSON is a PhD candidate at the University of Valencia. His research interests lie in the philosophy of language, with an emphasis on the conditions for successful communication.

ADDRESS: Valencia Philosophy Lab, Departament de Filosofia, Facultat de Filosofia i Ciències de l'Educació, Universitat de València, Av. Blasco Ibàñez 30, Valencia 46010, Spain. Email: valensimpson@gmail.com – ORCID:0009-0003-2263-8857

MATHEUS VALENTE is Assistant Professor at UNED, Spain. Who specializes in the philosophy of mind, language, and epistemology. He has published about indexical and self-locating attitudes, communication, reference, and demonstrative thoughts.

ADDRESS: UNED. Humanidades. Paseo de Senda del rey, 7 (28040 Madrid, Spain). Email: matheusvalenteleite@gmail.com – ORCID:0000-0001-6380-2623



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

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

Víctor M. Verdejo*

Pompeu Fabra University

<https://orcid.org/0000-0002-7872-0206>

Keywords

Bermúdez
Self
Frege's Criterion
Reference rule
Thought sharing

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

Palabras clave

Bermúdez,
Yo
Criterio de Frege
Regla de referencia
Compartir pensamientos

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

Gako-hitzak

Bermúdez
Nia
Fregeren irizpidea
Erreferentzia-araua
Pentsamenduak partekatzea

LABURPENA: Demagun zera esaten duzula: «Filosofo profesionala naiz». Adieraz al dezaket nik —edo beste norbaitek— zuk esaldi horrekin adierazi duzun pentsamendu bera? Baietz erantzunez gero, bereizi egin litezke adierazi duzun pentsamenduaren erreferentea eta beraren pentsalaria, eta aukera egon liteke pentsamendu hori lehen pertsonako izenordaina erabili gabe adierazteko. Hemen, intersubjektiboki parteka daitekeen pentsamendu mota hori (hots, «ni» motakoa) indibiduatu duen erreferentzia-arau baten oinarritzko ezaugarriak azaltzen ahaleginduko naiz. Horretarako, Bermúdezek (2019) proposatutako erronkari erantzun behar zaio; hots, hauek zehazteko erronkari: (a) zeri jarraikiz den banakoa erreferente baina ez nahitaez kasu horietako pentsalaria, eta (b) zergatik erabiltzen den esaldi bat «ni» motako pentsamendu bat adierazteko, eta ez beste mota bateko pentsamendu bat. Proposatuko dudana azalpena guztiz bat dator Bermúdezen beraren azalpenarekin kokapen-gaitasunei dagokienez. Hala ere, hark iradokitzen duenaren kontra, analisi, oraindik ere, testuinguruaren arabera mugatutako motaren (edo mota instantziagarriaren) mailan dago ziurrenik, eta horrek pentsamendua indibiduatzeko Fregeren irizpidea berrikustera —zuzenean baztertzera ez esatearren— garamatza.

* **Correspondence to:** Víctor M. Verdejo. Departament d'Humanitats, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain – victor.verdejo@upf.edu – <https://orcid.org/0000-0002-7872-0206>

How to cite: Verdejo, Víctor 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>).

Received: 08 May, 2024; Final version: 04 June, 2025.

ISSN 0495-4548 – eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

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

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

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

(i) I am a professional philosopher.

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

(ii) You are a professional philosopher.

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

These niceties turn out to lead to remarkable complexity in the satisfactory account of the type of thought involved, viz. the self type. Bermúdez's challenge concerns specifically those approaches which would wish to individuate this type —and indeed any type of thought— by means of an appropriate reference rule.⁴ According to Bermúdez (2019, 630–635), such an account might allow for shareability, as a reference rule for the self type can be devised to distinguish the referent (or as Bermúdez suggests, the subject) and the thinker of an instance of the thought. Nonetheless, the account

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

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

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

⁴ These are what Coliva & Palmira (2025) and others aptly call "metasemantic" accounts.

would be at fault without further determining (a) in virtue of what is an individual the referent but not necessarily the thinker of the self type of thought expressed with a non-first-personal sentence? To put it another way, what can enable a self type of thought to refer to an individual different from its thinker? And (b) what makes it the case that such a non-first-personal sentence would express a self type of thought, as opposed to a thought of a different type? Or, in other words, what are the reasons to suppose that a thought of precisely the self type is being expressed in these cases, as opposed to, say, a thought of a demonstrative or descriptive type?

I here advance an elucidation of the basic features of a token-reflexive reference rule individuating the shareable self type to specifically respond to Bermúdez's important challenge. In the course of doing so, I shall also examine the question of what is the correct level of analysis for the characterization of the self type and, relatedly, whether the resulting view is compatible with a baseline principle for individuating thought in terms of cognitive significance, namely, the so-called 'Frege's Criterion' or 'Frege's Constraint'. I shall contend that the account offered here is entirely congenial to Bermúdez's own account in terms of location abilities. However, contrary to what he suggests (Bermúdez 2019), the analysis plausibly remains one at the level of the contextually constrained type (or 'instantiable type'), and leads to a revision —if not a straightforward rejection— of Frege's Criterion for the individuation of thought.

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

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

(SYMMETRY)

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

(COGNITIVE)

Thoughts are individuated at the level of cognitive significance.

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

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

Arguably, only the instantiable type (i.e. the type as constrained by explanatorily salient instantiable features in the context) seems significant in discussions of the *interpersonal* shareability of thought expressible with the first person (Ver-

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

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

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

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

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

But now the challenge for Verdejo's account should be clear. It is all very well to define a version of the token-reflexive rule that applies to the subject of the thought, rather than to its thinker, but the key question now becomes: What makes an individual the subject of a first person thought in those cases where the subject of the thought is distinct from its thinker? [...] in virtue of what is A the subject of the first person thought that B expresses with (ii)? It is not hard to explain why A would be the subject of a *second* person thought that B might express with [(ii)]. But Verdejo needs to explain how (ii) can express a *first* person thought. Moreover, Verdejo also needs to explain what makes it the case that sentences such as (ii) express second person thoughts in some contexts, and first person thoughts in others. (Bermúdez 2019, 634-635)

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

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

3. Meeting Bermúdez's challenge

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

Much more could be said about the nature of subjects featuring in this reference rule, and the kinds of awareness of a subject sustaining instantiations of the self type. Various options would be open in the full elucidation of the condition "the subject" or "the person" of the thought in question. For instance, it is I think a common presupposition of Bermúdez's and other proposals that a subject here refers to a self-conscious individual or an individual that is able to self-refer (e.g., Bermúdez 1998, §1.4; Sainsbury 2011; Echeverri 2020, §2). Another key feature here seems to be that this subject is the producer of the thought in question (e.g., Campbell 1994, 73; O'Brien 2007, 49).⁸ To move the discussion forward, I will here remain neutral regarding these refinements. On any of these construals, however, the self rule leaves open whether or not the subject acting as the referent of the thought in the context is, moreover, the thinker of the thought in question.

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

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

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

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

⁸ This last suggestion, which emphasises the agentive aspect of the referent of the self type of thought, comports nicely with Bermúdez (2018) treatment of the sense of agency, but the argument in the text does not require it. See also Verdejo (2025) for relevant discussion.

by recourse to distinct first person or second person pronouns. From this angle, an utterance of (ii) in c expresses a self type of thought iff it involves the instantiation of a thought individuated by the self rule.

One may press on at this point and ask: what makes an utterance of (ii) referring to A be one that involves an application of the self rule as opposed to another rule individuating a co-referential thought in the context? For instance, why is that utterance not an application of, as it might be, the rule to refer to the salient professional philosopher in the context, or the rule to refer to one's addressee in the context? To appropriately answer this question we will need to attend to the particular explanatory contribution made by different proposals. First, the ascription of a shareable self type would need to be in harmony with the different phenomena traditionally associated with thought expressible with the first person pronoun, such as immunity to error through misidentification (Shoemaker 1968), and related properties such as guaranteed self-reference (Coliva 2003; Echeverri 2020; García-Carpintero 2024) or so-called 'illusions of transcendence' (Coliva 2012; Salje 2020). But there is no obvious reason to suppose that the self rule cannot accommodate these phenomena. In fact, for example, I have argued that the reference-rule articulation of the shareable self type may feature in explanations of immunity to error through misidentification that better capture its ground-relativity (Verdejo 2021b).

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

4. *The instantiable self type and location abilities*

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

Importantly, Bermúdez also maintains that sameness of thought in these cases involves two cognitively different ways of exercising an ability associated with (i) and (ii) respectively. In Bermúdez's words: "Of course, the way in which you go about determining whether I am F may be very different from how I go about determining whether I am F. The first-per-

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

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

son and second-person ways of establishing F-ness may differ, but that has no bearing on the relation between your understanding of “I” and my understanding of “I”, both of which exploit a single ability” (Bermúdez 2017, 116-117).

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

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

Bermúdez disagrees. He takes his analysis to be “firmly pitched at the level of token-sense” instead (Bermúdez 2019, 637). According to Bermúdez, the token-sense of ‘I’ is to be understood “in terms of abilities to determine the truth-condition of token sentences involving ‘I’ —and in particular the ability to locate the referent of the sentence” (ibid, 635). So, on Bermúdez’s view, the question of whether —in which contexts— do utterances of (i) and (ii) by A and B, respectively, involve the same thought or not is, in fact, the question of whether such utterances involve the same ability to determine the truth-condition of the corresponding sentences —and the location in space of A. But this way of putting things does not contravene the idea that token-sense is a notion at the level of instantiable type.

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

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

5. *Symmetry and Frege’s Criterion*

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

As noted, however, the present discussion also assumes COGNITIVE (sect. 2). The considered accounts of the (shareable) sense of ‘I’ thus factor in the distinctive cognitive significance that is associated with utterances of ‘I’ in the relevant cases. This means, for instance, that the sense expressible with ‘I’ must generally be distinguished from the sense expressible, say, with other potentially co-referring terms such as ‘that professional philosopher’ or ‘José’. Importantly, however, COGNITIVE involves no straightforward commitment to Frege’s Criterion. Put bluntly, Frege’s Criterion is a (well-known) prin-

ciple for spotting different thoughts in accordance with COGNITIVE, yet COGNITIVE could perfectly stand even if Frege's Criterion falls insofar as we have a positive story to tell about cognitive-significance based individuation. In fact, there is an initial motivation to suspect that, granted SYMMETRY, Frege's Criterion must indeed fall in the light of the positive stories already sketched above. For to accede to sameness of thoughts expressible with 'I' and 'you' is, as we have seen, also to accede to different (first- and second-person) ways of exercising an ability (on Bermúdez's account), or different thinker roles operative in thinking one and the same self type of thought (on the reference-rule account). The tension is plain if we take Frege's Criterion to discriminate types of thought in terms of such disparate ways or roles.

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

Frege's Criterion

For any thoughts, T and T' , and subject S , if S can rationally believe (be agnostic about/reject) T and be agnostic about/reject (believe) T' , then $T \neq T'$.

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

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

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

Frege's Criterion in Context

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

This formulation involves a relativization to sentences and contexts of utterance—or "utterance-tokens" as Bermúdez puts it. Still, the relevant question is a possibility question, namely: is it really *possible* for a thinker to rationally take different attitudes to co-referring¹³ thoughts (or token-senses) expressible with utterances of 'I' and 'you' (as in (i) and (ii) above) in a context?

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

¹² See Evans 1982; Harcourt 1999, 331. Evans quotes Frege's letters to Russell and Jourdain (1980, 80 and 153) in support of a similar reading of the criterion. It is logically entailed by the stronger equipollence criterion for the sameness of thought (Frege 1979, 197-198), also alluded to by Evans.

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

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

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

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

Thus, the fact that a subject in a context may not be allowed to rationally have conflicting attitudes towards the thought expressed by a pair of sentences—such as sentences of the form 'Hesperus is F' and 'Phosphorus is F' as understood by someone who knows that Hesperus is Phosphorus—is orthogonal to the question of whether the thoughts so expressed are distinct or not according to the criterion. For the same reason, the fact that there is a context in which one is not allowed to have conflicting attitudes towards 'I'/'you'-thoughts of the sort expressed by (i) and (ii)—the contexts targeted by Bermúdez's Symmetry Constraint—is orthogonal to the question of whether Frege's Criterion is ultimately respected.

More may be ventured, for to grant—as Bermúdez does in the quoted passage above (Bermúdez 2019, 637)—that 'I'/'you'-thoughts as expressed with (i) and (ii) will cognitively differ in some contexts and not others, already indicates that Frege's Criterion must prescribe a difference in the thoughts so expressed. We can confirm this possibility with many examples. For instance, we may conceive of David Kaplan, John Perry or certainly ourselves accepting the thought expressed with (i) while at the same time squarely rejecting or being agnostic about the thought expressed with (ii) when inadvertently referring with 'you' to themselves as appearing on a blurry TV screen.

6. Symmetry and double contextualization

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

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

The compatibilist objection would not here exactly be levelled against the idea that Frege's Criterion ought to be articulated in such a way that sameness of thought matches sameness of type of expression. For, in order for 'I'/'you'-thought

sameness to make sense at all —and assuming that 'I' and 'you' are indeed different types of expression— it already has to be true that thought-type identity doesn't require expression-type identity (see also sect. 3). The compatibilist invokes rather an interpretation of Frege's Criterion such that sameness of reference for one and the same expression type (viz. the linguistic types corresponding to 'I' or 'you') does not necessarily yield sameness of thought type across contexts.

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

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

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

Frege's Criterion with Double Contextualisation

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

Now, while there is nothing incoherent in this strategy per se, it does little by way of preserving Frege's Criterion as a substantive principle for the individuation of thought. Once we concede that the cognitive differences registered by means of the criterion may have an effect in some contexts and not others —and do so irrespective of co-reference, sameness of expression type or standard of language use (for 'I' and 'you')— we open the door to any thought-difference detected by means of the criterion being arguably irrelevant for the context at stake, and indeed for any context other than the context in which the difference is detected (that is, irrelevant whenever $c_A \neq c_I$).¹⁵

¹⁴ This is also a reading of Frege's Criterion that can be endorsed to respond to what Valente calls "non-transitivity cases" or cases that purportedly show that Symmetry is incompatible with the intransitivity of communication or testimonial knowledge transmission (Valente 2025, §2; see also Valente & Onofri 2023; Goodman 2025). For reasons made clear in the main text, I agree with Valente that this kind of contextual relativisation is counterproductive.

¹⁵ I take this point to apply, *mutatis mutandis*, to the articulation of Frege's Criterion in terms of actual, as opposed to absolute, possibility advocated by Recanati (2025). Detailed discussion of Recanati's views in this connection must be left for another occasion.

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

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

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

In consequence, it is only at t_1 that we may have the right to claim that thoughts expressible with ‘Hesperus’ and ‘Phosphorus’ are guaranteed to be distinct according to the criterion. But to concede such stark contextualisation backfires into idleness. For if Hammurabi becomes aware of the fact that ‘Hesperus’ and ‘Phosphorus’ co-refer at a later time t_2 it would cease to be true, according to the criterion so understood, that ‘Hesperus’/‘Phosphorus’-thoughts are distinct. For at t_2 , and contrary to Evans’s insight highlighted above, thought distinctness will not persist if —as a consequence of newly acquired knowledge— it ceases to be rational for Hammurabi to stand in conflicting attitudes towards thoughts expressed with ‘Hesperus’ and ‘Phosphorus’. In sum, if we restrict difference of thought to t_1 , the criterion turns out to be idle as a principle of individuation: it would at most reflect the specificities of t_1 but not the defining features of ‘Hesperus’/‘Phosphorus’-thoughts as such. In general, if we restrict the context of individuation to the context of application, even paradigmatic distinctions delivered by the criterion in non-indexical cases would run the risk of being idle for the individuation of thought.

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

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

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

7. *Going full circle: instantiable types again*

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

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

Acknowledgements

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

REFERENCES

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

- Bermúdez, J. L. (2019). First person thoughts: Shareability and symmetry. *Grazer Philosophische Studien*, 96(4), 629-638.
- Campbell, J. (1994). *Past, space, and self*. Oxford: Clarendon Press.
- Chalmers, D. (2011). Propositions and attitude ascriptions: A Fregean account. *Noûs*, 45(4), 595-639.
- Coliva, A. (2003). The first person: Error through misidentification, the split between speaker's and semantic reference, and the real guarantee. *Journal of Philosophy*, 100(8), 416-431.
- Coliva, A. (2012). Which 'key to all mythologies' about the self? A note on where the illusions of transcendence come from and how to resist them. In S. Prosser & F. Recanati (Eds.). *Immunity to error through misidentification: New essays* (pp. 22-45). Cambridge: Cambridge University Press.
- Coliva, A. & Palmira, M. (2025). Immunity to error through misidentification: some trends. *Philosophical Psychology*, 38(3), 1126-1161. <https://doi.org/10.1080/09515089.2024.2371423>
- Davidson, D. (1970). Mental events. Reprinted in D. Davidson (Ed.). *Essays on action and events* (pp. 207-224). Oxford: Clarendon Press, 1980.
- Echeverri, S. (2020). Guarantee and reflexivity. *Journal of Philosophy*, 117(9), 473-500.
- Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.
- Frege, G. (1956). The thought: A logical inquiry. Translated by A.M. and M. Quinto. *Mind*, 65(259), 289-311.
- Frege, G. (1979). *Posthumous writings*. Translated by P. Long and R. White. Chicago: University of Chicago Press.
- Frege, G. (1980). *Philosophical and mathematical correspondence*. Abridged from the German edition by Brian McGuinness. Translated by Hans Kaal. Oxford: Basil Blackwell.
- García-Carpintero, M. (2024). The *real guarantee* in *de se* thought: How to characterize it? *Philosophical Quarterly*, pqae133. <https://doi.org/10.1093/pq/pqae133>
- Goodman, R. (2025). Shared thought and communication. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 218-239). Oxford University Press.
- Harcourt, E. (1999). Frege on "I", "now", "today" and some other linguistic devices. *Synthese*, 121(3), 329-356.
- Howell, R.J. (2023). *Self-awareness and the elusive subject*. Oxford: Oxford University Press.
- Kripke, S. (2011). The first person. In S. Kripke (Ed.). *Philosophical troubles, volume 1* (pp. 292-321). New York: Oxford University Press.
- Longworth, G. (2013). Sharing thoughts about oneself. *Proceedings of the Aristotelian Society*, 113(1), 57-81.
- Longworth, G. (2014). You and me. *Philosophical Explorations*, 17(3), 289-303.
- May, R. (2006). Frege on indexicals. *Philosophical Review*, 115(4), 487-516.
- O'Brien, L. (2007). *Self-knowing agents*. Oxford: Oxford University Press.
- Recanati, F. (2025). Sameness of mode of presentation. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 193-217). Oxford University Press.
- Sainsbury, M. (2011). English speakers should use 'I' to refer to themselves. In A. Hatzimoysis (Ed.). *Self-knowledge* (pp. 246-260). New York: Oxford University Press.

- Salje, L. (2020). Lit from within: First-person thought and illusions of transcendence. *Canadian Journal of Philosophy*, 50(6), 735-749.
- Schellenberg, S. (2012). Sameness of Fregean sense. *Synthese*, 189(1), 163-175.
- Shoemaker, S. (1968). Self-reference and self-awareness. *Journal of Philosophy*, 65(19), 555-567.
- Skiba, L. (2017). Frege's unthinkable thoughts. *Proceedings of the Aristotelian Society*, 117(3), 333-343.
- Soames, S. (2015) *Rethinking language, mind, and meaning*. Princeton (NJ): Princeton University Press.
- Speaks, J. (2013). Individuating Fregean sense. *Canadian Journal of Philosophy*, 43(5), 634-654.
- Textor, M. (2018). Frege's recognition criterion for thoughts and its problems. *Synthese*, 195(6), 2677-2696.
- Travis, C. (2013). *Perception: Essays after Frege*. Oxford: Oxford University Press.
- Valente, M. (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>
- Valente, M. & Onofri, A. (2023). A puzzle about communication. *Review of Philosophy and Psychology*, 14(3), 1035-1054.
- Verdejo, V.M. (2018a). On having the same first person thought. *Grazer Philosophische Studien*, 95(4), 566-587.
- Verdejo, V.M. (2018b). Thought, communication and perspectives about the self. *Dialectica*, 72(4), 487-507.
- Verdejo, V.M. (2019). In defence of the shareability of Fregean self-thought. *Acta Analytica*, 34(3), 281-299.
- Verdejo, V.M. (2021a). The second person perspective. *Erkenntnis*, 86(6), 1693-1711.
- Verdejo, V.M. (2021b). Perspectives on *de se* immunity. *Synthese*, 198(10), 10089-10107.
- Verdejo, V.M. (2025). On the rationality of thought insertion judgments. *Philosophical Psychology*, 38(5), 2076-2099. <https://doi.org/10.1080/09515089.2023.2296580>

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

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



ASYMMETRIES BETWEEN ‘YOU’ AND ‘I’

(*Asimetrías entre ‘tú’ y ‘yo’*)

Matheus Valente*

UNED

<https://orcid.org/0000-0001-6380-2623>

Keywords

Thoughts
First-person
Propositional attitudes
Indexicals
Action

ABSTRACT: Can others grasp my first-person thoughts, or are such thoughts inherently private? Philosophers disagree: some argue that first-person thoughts are apprehensible only by their owners, while others contend that they can be shared through communication—expressible by ‘you’ as readily as by ‘I’. In this paper, I set out to clarify the stakes of this age-long dispute. Taking J. L. Bermúdez’s forceful defence of shareability as the backdrop of my discussion, I examine how the intersubjective availability of thoughts interacts with issues concerning the objectivity of thought, testimonial knowledge transmission, and rational action. The bulk of this paper is an elaboration of the Asymmetry Argument, which grounds the privacy of first-person thoughts in the need to explain how thinkers who believe and desire the same as each other might nonetheless have distinct reasons for action. If successful, the argument reveals how first-person thoughts cannot be shareable in a philosophically significant sense without compromising their fundamental connection to motivating reasons for action.

Palabras clave

Pensamiento
Primera persona
Actitudes proposicionales
Deícticos
Acción

RESUMEN: ¿Pueden otras personas comprender mis pensamientos de primera persona, o son estos pensamientos inherentemente privados? En filosofía no hay consenso: una corriente defiende que los pensamientos de primera persona sólo son aprehensibles por sus poseedores, mientras que otra sostiene que pueden compartirse mediante la comunicación – siendo expresables tanto por «tú» como por «yo». En este artículo, me propongo aclarar los puntos clave de esta antigua disputa. Partiendo de la contundente defensa de la compartibilidad de los pensamientos de primera persona realizada por J. L. Bermúdez, examino cómo la disponibilidad intersubjetiva de esos pensamientos interactúa con cuestiones relativas a la objetividad del pensamiento, la transmisión del conocimiento por testimonio y la acción racional. El núcleo de este artículo es una elaboración del Argumento de la Asimetría, que fundamenta la privacidad de los pensamientos de primera persona en la necesidad de explicar cómo pensadores que creen y desean lo mismo podrían, no obstante, tener distintas razones para actuar. De tener éxito, el argumento revela cómo los pensamientos de primera persona no pueden considerarse compartibles sin comprometer su conexión fundamental con las razones para actuar.

Gako-hitzak

Pentsamenduak
Lehen pertsona
Jarrera proposizionalak
Deiktikoak
Ekintza

LABURPENA: Uler al ditzakete besteek lehen pertsonako nire pentsamenduak, edo pribatuak dira berez? Filosofoak ez datoz bat: batzuek argudiatzen dute lehen pertsonako pentsamenduak dauzkatenek bakarrik atzeman ditzaketela, eta beste batzuek, berriz, komunikazioaren bitartez parteka daitezkeela —eta «ni» gisa bezain erraz adieraz daitezkeela «zu» gisa—. Lan honetan, saiatuko naiz argitzen zer dagoen jokoan aspaldiko gatazka honetan. J.L. Bermúdezek lehen pertsonako pentsamenduen partekagarritasunaz egiten duen defentsa irmoa nire analisiaren oinarri gisa hartuta, arretaz aztertuko dut nola eragiten dioten elkarri, batetik, pentsamendu horien eskuragarritasun intersubjektiboak eta, bestetik, pentsamenduaren objektibotasunari, testigantza bidezko ezagutzaren transmisioari eta ekintza arrazionalari lotutako auziek. Lan honen zatirik handienean, Asimetriaren Argudioa landuko dut. Artikulu honen helburua da asimetriaren argudioa garatzea, lehen pertsonako pentsamenduen pribatutasuna azaldu behar honetan oinarrituz: nolatan gauza bera uste edo nahi duten pentsalariek halere ekintzarako arrazoi desberdinak izan ditzaketen. Arrakasta izanez gero, argudioak agerian utziko du lehen pertsonako pentsamenduak ezin direla partekatu filosofikoki esanguratsua den zentzu batean, ekintzara daramaten arrazoiekiko beren funtsezko lotura arriskuan jarri gabe.

* **Correspondence to:** Matheus Valente. UNED. Humanidades. Paseo de Senda del rey 7, 28040 Madrid, Spain. – matheusvalenteite@gmail.com – <https://orcid.org/0000-0001-6380-2623>

How to cite: Valente, Matheus (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>).

Received: 12 June, 2024; Final version: 02 June, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

When confronted with the peculiar features of first-person thought (those usually expressed with ‘I’), many philosophers have felt compelled to hold that they differ from third-person thoughts in being apprehensible only by their owners.¹ Frege (1956) himself wrote that “everyone is presented to himself in a special and primitive way, in which he is presented to no one else”, a remark often cited in defences of the thesis that first-person thoughts are private (henceforth, **Privacy**). Unsurprisingly, **Privacy** has not gone wholly unchallenged, even by authors of a broadly Fregean persuasion, such as José Luis Bermúdez.²

In a series of works spanning decades, Bermúdez has developed a systematic account of first-person thoughts among whose original features is a forceful defence of their objectivity and shareability. More specifically, Bermúdez argues for the so-called symmetry constraint - henceforth, **Symmetry** - which holds that “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” (2017a, p. 61). The idea behind **Symmetry** is as simple as it is plausible: when you tell me something about yourself with ‘I’, I acquire knowledge which I could report back to you with ‘you’, suggesting that reciprocal uses of the first- and second-person pronouns not only might happen to co-refer but also to match in cognitive significance, and so, to underlie the apprehension of a single thought.³

Despite extensive debate, the stakes of this dispute remain unclear. Some of this hinges on which theoretical roles thoughts are supposed to play. While most agree that thoughts are the cognitively significant contents of attitudes such as belief, it’s unclear whether this alone suffices to adjudicate between **Symmetry** and **Privacy**. This paper aims to achieve more clarity on the significance of this dispute.

Sections 2 and 3 assess Bermúdez’s two main positive arguments for **Symmetry**: *the argument from objectivity* and *the argument from testimonial knowledge*.⁴ I argue that neither succeeds; indeed, the latter one might actually backfire, indirectly supporting **Privacy** due to issues with the assumption that thought sharing is necessary for the transmission of knowledge through communication. Section 4 turns to the interplay between thoughts and motivating reasons for action, introducing the concept of *confidants* to capture the relationship between ideally rational agents whose beliefs and desires are maximally shared through an idealised communicative exchange.⁵ This sets the stage for the Asymmetry Argument (Section 5), which concludes that **Privacy** alone can explain how such agents might nonetheless have distinct reasons for action. After rebutting objections to the argument (Section 6), I conclude that the need to explain how agents who maximally agree with each other might diverge in their reasons for action gives rise to a direct argument for the privacy of first-person thoughts.

¹ Authors who subscribe to the privacy of first-person thoughts include Evans (1981, 1982), Heck (2002), Morgan (2009), and Stanley (2011).

² The main works where Bermúdez discusses first-person thoughts and the sense of ‘I’ are: 2005, 2011, 2017a, 2017b, 2019. All Bermúdez’s citations in this paper come from his monograph *Understanding “I”* (2017a), where his main ideas on the topic are given its most systematic presentation.

³ By ‘the same thought’ I mean ‘thought-tokens of the same type’. In Bermúdez’s terminology, these would be thought-tokens constituted by the same token-senses (Bermúdez, p. 80-97).

⁴ A disclaimer: I discuss only a small fraction of Bermúdez’s rich account of the sense of ‘I’. Among its many invaluable contributions, we find: a defence of **Essential Indexicality** (roughly, the thesis that first-person attitudes must necessarily be involved in explanations of actions; Bermúdez, 2017a, chapter 1; 2017b), a careful reevaluation of the legacy of Evans’ seminal work on the first-person (Bermúdez, 2017a, chapter 6), and an original account of immunity to error *via* misidentification and past-tense judgements based on autobiographical memory (Bermúdez, 2017a, chapter 7; 2011).

⁵ The interplay between first-person thoughts and action has figured prominently in recent work. For a survey, see Ninan (2016; 2021). See also: Prosser (2005; 2023; 2025), Longworth (2013, 2014), Weber (2014), Recanati (2016; 2025), Torre (2018), Valente (2018), Torre & Weber (2021; 2022), Gray (2022), Lin (2022), Verdejo (2017; 2019; 2020; 2026).

2. *The Argument from Objectivity*

Though deeply influenced by Gareth Evans (1981, 1982), Bermúdez departs from him on a few important issues, including on the shareability of first-person thoughts. One source of their disagreement centres on the objectivity of thoughts, a central Fregean commitment which both authors endorse. While Evans argues that private thoughts can be objective, Bermúdez insists that first-person thoughts can only be objective if shareable.⁶

What is it for a thought to be objective? Most generally, a thought is objective if it is irreducible to the contents of a particular consciousness (i.e. distinct from, as Frege would put it, mere ideas [*Vorstellungen*]). Evans and Bermúdez seem to agree that a thought is so irreducible when it fulfils at least one of the following conditions:

- (1) The thought exists and has a truth-value independently of being apprehended;
- (2) The thought is shareable (apprehensible by multiple thinkers);

Evans defended **Privacy** by appealing to (1): “an unshareable [private] thought can be perfectly objective —can exist and have a truth-value independently of anyone’s entertaining it” (Evans, 1981, p. 313). Bermúdez disagrees, arguing that indexical thoughts (of which first-person thoughts are a key instance) are constitutively connected to the episodes of thinking whereby they are apprehended (p. 62-66):

[T]he whole point of token-reflexivity, and indexicality in general, is that the identity of the thought is determined by the context in which it is thought —in which case, there is no thought without an episode of thinking. (p. 63)

On that basis, Bermúdez concludes that indexical thoughts can only be objective if they fulfil (2), i.e. if they are shareable. But why should the identity of indexical thoughts be constrained in the way which Bermúdez outlines? What follows is my interpretation of his reasoning.

Suppose I thought ‘I am tired’ last Sunday at noon.⁷ The thought I thereby apprehend —call it ‘T’— is referentially equivalent to (the thought which I could have expressed as) ‘MV was tired on Sunday at noon’. However, their cognitive significance differs. For example, I could know that I’m tired without knowing that I’m MV or what time it is, a familiar point which shows that my first-person thought T is potentially different from its third-person analogue. One reason why that is so stems from the independent ways in which these thoughts’ truth-conditions are fixed. In particular, T’s truth-conditions are fixed by contextual features of its episode of apprehension such as the fact that its thinker is MV and that it’s entertained on a particular Sunday at noon. Presumably, this entails that T would not so much as exist if it had not been apprehended in a context with these particular features, which plausibly tells against Evans’ contention that its objectivity could be grounded on fulfilment of (1).

However, we immediately face problems when trying to make Bermúdez’s proposal more precise. Surely, Bermúdez does not mean to say that T could not possibly have been apprehended in any other context. Suppose, for example, that I knew on Saturday 8pm that I would be tired on Sunday at noon, and thought ‘I will be tired in 16 hours’. Presumably, the thought I would then apprehend could be none other than T, after all, if I went on to track the passage of time for 16 hours and, finding myself tired just as I knew I would, thought ‘I am tired’, I would seem to be merely re-expressing the content of my earlier judgement as opposed to judging something new. To be sure, there are some who would challenge that intuitive claim, but these would be people who balk at the idea that an indexical thought could remain invari-

⁶ Frege’s emphasis on the objectivity of thoughts is arguably one of his central philosophical contributions (Dummett, 1980).

⁷ For brevity, I’ll often write “x thinks/believes ‘I am F’” as short for “x apprehends/believes a thought which they could have expressed by uttering ‘I am F’”.

ant across changes of perspective, and so, whose fundamental commitments are at odds with the core rationale behind **Symmetry**, i.e. that a thought expressible in the first-person perspective with 'I' could equally be expressed second-personally with 'you'. But if T could have been apprehended on Saturday, then it could have existed and have a truth-value independently of being apprehended on Sunday, which shows that its purported constitutive connection to a particular episode of thinking is less clearcut than it initially might seem.

Bermúdez could insist that T's existence depends on its being apprehended in *some* context, though not on any particular one. Notice, however, that some of these considerations appear to pertain more closely to the temporal indexicality of T as opposed to its involvement of the first-person sense of 'I'. I take it to be uncontroversial that a thought can be indexical in at least two ways: by involving the senses of personal pronouns like 'I' and 'you' or the senses of temporal indexicals like 'now' and 'soon'. T appears to be indexical twice over: not only it concerns myself as myself, it also makes indexical reference to time. In contrast, the thoughts 'I am tired at 12:00 GMT on Sunday, January 3rd, 2025' and 'today is Sunday' seem to be indexical in a single sense: the first is first-personal but not temporal, the latter is temporal but not personal. This raises a concern: in focusing on T to investigate the objectivity of first-person thoughts, we risk equivocating between issues which arise from temporal indexicality with possibly independent issues having to do with the first-person.⁸ To avoid any mixup, we should turn attention to pure breed first-person thoughts, of which these are plausible examples:

On Sunday, MV thinks: (T2) I am tired at 12:00 GMT on Sunday, January 3rd, 2025.

On Sunday, MV thinks: (T3) I am tired at some/all times.

On Sunday, MV thinks: (T4) I am human *simpliciter*.⁹

As far as I can see, Bermúdez's criticism of Evans is harder to motivate on the basis of examples like T2, T3, and T4. Why, for example, should the identity of T2 —the thought whereby I self-attribute tiredness at a particular, non-indexically specified, time - be in any way constrained by when or where it is apprehended? Perhaps Bermúdez could insist that these thoughts' truth-conditions are fixed by an indexical act of reference to a particular individual— either via 'I' or, as **Symmetry** allows, via 'you' —which is subsidiary to an episode of apprehension. I agree that that something gets to be the referent of a token of 'I' or 'you' only if these tokens were indeed produced in the right way, but that strikes me as a trivial claim equivalent to saying that a token singular term ought to exist in order for it to have a referent. The crucial question is whether it makes sense to say that thoughts like T2, T3, and T4, could have existed and have a truth-value even if nobody had ever entertained them, and so far I have not been able to pinpoint a clear reason for saying that it does not. On the other hand, a reasoning like the following moves me into saying that it does make sense.

Suppose that I have known all along of this person who I fail to recognise as myself that she is tired at 12:00 GMT on Sunday, January 3rd, 2025. At some point in February, I realise that this person is none other than myself, and so, discover that it is me who was tired at 12:00 GMT on Sunday, January 3rd, 2025, thus apprehending T2 for the first time around. Intuitively, the truth which I discovered, T2, 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?

Furthermore, in the scenario under consideration, I have apprehended all of the senses constitutive of T2 —the objectual sense of 'I' and the sense(s) of 'being tired at 12:00 GMT on Sunday, January 3rd, 2025'— prior to concatenating

⁸ For a recent work addressing the peculiar differences between personal (*de se*) and temporal (*de nunc*) indexicality, see Morgan (2024).

⁹ Whenever F is an essential property such as being human, I assume that it makes sense to think of an individual as being F *simpliciter* (as opposed to: being F *now/in the past/in the future*).

them as I do when I finally apprehend T2, after all, I have thought about myself and, independently, about someone who was tired on that specific date. Presumably, this entails that the constituents of T2 were available for thought prior to my apprehending it. If so, then Bermúdez seems committed to denying that a thought exists and has a truth-value even though all of its constituents have been separately apprehended in a meaningful way. But why should a thought's existence have to wait any longer?

As said before, the objectivity of a thought most fundamentally requires its being irreducible to mere ideas. Plausibly, an idea's subjectivity has at least partly to do with its fleeting, ephemeral nature; but thoughts like T2, T3, and T4, certainly can be entertained multiple times (even if only by a single subject), and so, wouldn't seem to be fleeting in a similar way. Interestingly, this fits with another traditional conception of objectivity, namely, that of existence unperceived. For example, one might reason by analogy and claim that, just like an objective particular is one which may be re-encountered after ceasing to be perceived, an objective thought is one which may be re-apprehended after previously having been so. Why, then, would objectivity require first-person thoughts to be apprehensible by multiple thinkers over and above requiring them to be apprehensible multiple times by a single one?

Rather than a set of knockdown objections, the above should be read as a plea for further clarification regarding the rules of this debate. As things stand, I'm not convinced that **Symmetry** and the shareability of first-person thoughts can be vindicated solely on the basis of their objectivity.

3. *The Argument from Testimonial Knowledge*

Bermúdez bases his other main argument for **Symmetry** on another broadly Fregean principle, namely, that “a Fregean theory of sense is a theory of understanding” (p. 67), and so, that there ought to be an “equation between understanding a sentence and grasping the thought that it expresses” (p. 27). On a natural interpretation, this requires thinkers who understand each other to apprehend the same thought. Bermúdez (p. 76) is careful to accept that not all forms of understanding might require thought sharing, but he explicitly claims that this must be so when knowledge gets transmitted via testimony. Thus, Bermúdez's main claim is that “knowledge through testimony typically requires the person acquiring the knowledge to think the very thought expressed by the speaker” (p. 76). If he's right, then **Symmetry** could be supported by the undeniable fact that one may acquire knowledge from another's testimony involving ‘I’, knowledge which one would naturally report back with ‘you’.

However, Bermúdez's claim faces at least two serious problems. First, the relevant data on testimonial knowledge transmission appears amenable to an account which deploys no notion more sophisticated than referential contents and independently motivated epistemic concepts such as safety and reliability. Second, so-called *non-transitivity cases* appear to show that knowledge transmission can occur even when thought sameness is absent. I now explain each of these issues.

A quick look at the recent literature on referential communication shows that the following is common ground among authors of virtually all persuasions: sameness of referential content is not sufficient for understanding and knowledge transmission via testimony.¹⁰ For example, if I tell you ‘Clark Kent is F’ and you form the belief ‘Superman is F’, then, assuming that neither of us know that Clark Kent is Superman, your belief would have the same referential content as mine but wouldn't count as knowledge. A Fregean might say that your mistake was failing to apprehend the same thought as mine. But a non-Fregean —call her a Russellian— will judge this move *ad hoc*, for we seem capable of explaining your failure in purely referential-*cum*-epistemic terms. For example, “the Russellian might claim that the process involved in communication [knowledge transmission] must reliably produce coreferential thoughts, and so secure

¹⁰ Cases with this structure are often called ‘Loar-cases’ due to their origin in Loar (1976). For recent debate involving such cases, see Onofri (2019), Valente (2021), Prosser (2025), Recanati (2016, 2025).

non-accidental, or not merely lucky, co-reference" (Goodman, 2025, p. 221). In other words, you failed because you got the right referential content by sheer luck as opposed to in a suitably reliable way. As far as I can see, there's nothing immediately unobjectionable about that kind of story (not to mention the fact that it's more theoretically parsimonious, after all, it can be endorsed independently of any prior commitments regarding the role of thoughts in interpersonal exchanges).

A Fregean might retort that what the Russellian calls 'apprehending the right referential content in a reliable way' just is what she calls 'apprehending the same thought'. But there are problems with this response. For example, some have argued that thought sameness can itself be arrived at as a matter of luck (Byrne & Thau, 1996). Suppose that, by sheer coincidence, you and I independently introduce the name 'Winston' for the amnesiac in Room 101. I then tell you, 'Winston will never recover' and you form the belief you'd express by the same sentence. Plausibly, your belief won't count as knowledge: how do you know who I'm talking about? However, it's possible to argue that the thought which you believe is the same as that which I have expressed, after all, the senses you and I associate to our uses of 'Winston' appear to be the same, namely, a sense for the amnesiac in Room 101. If so, then even Fregeans will need to appeal to anti-luck considerations in order to characterise knowledge-supporting understanding, which again suggests that we might do better by following the Russellian in appealing to nothing but referential-*cum*-epistemic terms to accommodate the relevant phenomenon.

To respond, a Fregean needs to come up with a principled argument for the importance of having a thought-involving account of the relevant data. Let's assume that this can be done. Still, it's not obvious whether this would result in any direct arguments for **Symmetry**. For one thing, knowledge-supporting understanding might require some suitable relation between the relevant thoughts without requiring them to be the same. Indeed, some have advanced proposals in that spirit. McDowell (1984, p. 290) is a good example:

[...] there is no obvious reason why [Frege] could not have held [...] that in linguistic interchange of the appropriate kind, mutual understanding—which is what successful communication achieves—requires not shared thoughts but different thoughts which, however, stand and are mutually known to stand in a suitable relation of correspondence.

Therefore, even Fregeans can eschew the need for thought sameness as a condition on testimonial knowledge transmission. But why would any Fregean go down that route instead of sticking to the simpler view endorsed by Bermúdez? There might be multiple reasons why, but I'll resort myself to presenting that which I take to be the most general one: non-transitivity cases (see below) show that thought sameness not only isn't sufficient for knowledge transmission, it's not even necessary.¹¹

Suppose that I meet Johannes on Monday and tell him that I'm human. I then meet him again on Tuesday and, again, tell him I'm human. It's certainly possible for Johannes to understand me both times, and so, to come to know all that I told him. Presumably, the thought I express to him is the same both times (T4 from Section 2). So, if Bermúdez's view is correct, then what Johannes apprehends and comes to know on Monday and Tuesday is thought T4. However, Johannes might have failed to recognise me on Tuesday as the same person who talked to him on Monday. Surely, this type of identification failure happens in real life all the time. But if he did, then the thoughts he apprehends on each day are cognitively distinct for him, i.e. he could in principle believe one while disbelieving the other (and rationally so). Of course, this just is to say that Johannes apprehends different thoughts each time, contradicting Bermúdez's assumption that knowledge transmission requires thought sameness.

To be sure, I admit that non-transitivity cases are anything but unobjectionable, e.g. one might suspect that Johannes' recognition failure tells against his coming to know what I tell him on Tuesday. I agree that recognition failures might

¹¹ Non-transitivity and thought sharing has been a focus of recent debate by, among others, Goodman (2025); Gray (2025); Valente & Onofri (2023); Valente & Verdejo (2021), Cumming (2013), Fine (2007).

sometimes lead to misunderstanding, but doubt that this needs always occur. It would be a stretch to claim that I cannot acquire knowledge from your utterance ‘NN is F’ if I fail to see that NN is the same person you were talking about some time before. As far as I can see, this is analogous to how I could look at an object and learn that it is red while later, failing to recognise the same object as before, look at the same object and learn that it is red, ending up with two pieces of knowledge that are based on similar perceptual experiences of the same thing.¹² So, if one agrees that Johannes could understand me regardless of not recognising me, then he can only come to know what I tell him on Tuesday if he apprehends a thought that is different from mine. In addition, one might claim that Johannes would have failed to understand me on Tuesday if he apprehended the same thought as mine, for then he would have been assuming that I’m the guy from yesterday in the absence of good reasons for doing so.

How would Bermúdez respond? In a couple of places, he suggests an argument as follows (cf. Bermúdez, pp. 75-77). Suppose a bystander eavesdropped on my Tuesday conversation with Johannes. If the bystander understood my ‘I am F’ as well as Johannes’ ensuing report ‘you are F’, then it would be irrational for her to take contrasting attitudes to what Johannes and I have said. Doesn’t this show that Johannes and I express the same thought by ‘you’ and ‘I’ on that occasion? I don’t think so. This at most shows that the bystander couldn’t have understood us if she associated distinct thoughts to each of our utterances. This puts demands on the bystander’s thought(s) while leaving it open whether it/they ought to be same as mine, Johannes’, or to both. More generally, rational cotenability is a relation between the thoughts of a single thinker at a time, and so, are bound to be of little help in assessing distinct thinkers’ thoughts for identity or distinctness (Evans 1981).¹³

In conclusion, even if we grant the assumption that an account of knowledge transmission via testimony must be framed in terms of thinkers’ thoughts, thought sameness might still neither be sufficient nor necessary for it to take place. For these reasons, I’m sceptical about whether any straightforward argument from knowledge transmission to **Symmetry** will be forthcoming. Given this negative conclusion, we’re advised to look elsewhere in order to adjudicate between **Symmetry** and **Privacy**. In the next section, we turn to motivating reasons for action.

4. *Attitudes, Reasons, and Confidants*

As Bermúdez reminds us, it’s a “basic tenet of propositional attitude psychology that people act the way they do because of what they believe” (p. 47). Indeed, it’s almost a truism that a rational agent who believes that Φ ing is the optimal way of fulfilling her most pressing desires, ideally, will be motivated to Φ (and so, will predictably try to Φ). This gives rise to a sort of platitude about ideally rational agents, namely, that their motivating reasons for action —those in light of which she intentionally acts (henceforth, her ‘reasons’)¹⁴— supervene on her attitudes (e.g. beliefs, desires, etc). This supervenience claim can be given an intra-personal diachronic formulation and an inter-personal synchronic one:

¹² Furthermore, the fact that these cases involve multiple contexts/conversations strikes me as irrelevant, for intra-contextual analogue cases could easily be constructed (e.g. if ‘NN’ is a very common name, you might wrongly but rationally take my series of utterances about NN in a conversation as being about distinct people with the same name; for related considerations, see Verdejo, 2026).

¹³ This draws from Evans (1981, pp. 293-5), but see also Recanati (2025) and Verdejo (2026). An anonymous referee points out further issues with the intuitive Fregean criterion for thought individuation, in particular, a version thereof which takes thoughts to be the same when contrasting attitudes towards them are not rationally cotenable. As the referee points out, this criterion over generates, for it seemingly entails that analytic or a priori statements express the same thought (e.g. “all planets revolve” and “all bachelors are unmarried”). This seems right. In addition, I may add, it has the unfortunate consequence of entailing that co-referential thoughts known to be co-referential are the same (e.g. if I know that I am NN, then I express the same thoughts by ‘I am F’ and ‘NN is F’).

¹⁴ Motivating reasons, i.e. reasons in light of which an agent acts, are typically contrasted with normative reasons, i.e. reasons that count in favour of actions independently of whether the agent is aware or acts in their light (Alvarez & Way, 2024). My discussion primarily concerns the former. For recent discussion on motivating reasons and similar proposals regarding their connection to agents’ attitudes, see Singh (2019) and Wang (2025).

CHANGES: an ideally rational agent’s reasons cannot change if her attitudes remain the same.

TWINS: ideally rational agents cannot have distinct reasons if they have all the same attitudes.¹⁵

CHANGES says of an agent at distinct times what **TWINS** says of distinct ones at the same time. Presumably, both principles stand or fall together. This becomes even more crystalline if we reframe **CHANGES** as the principle that time-slices of an ideally rational agent cannot have different reasons if they have the same attitudes. Absent reasons to think that the relation between the attitudes and reasons of time-slices of an agent differ from those of time-slices of distinct agents, **CHANGES** and **TWINS** just are two faces of the same coin.

Since **Symmetry** is an inter-personal principle, let’s focus on **TWINS** (I’ll have more to say about **CHANGES** in 6.1). I take it to be uncontroversial that **TWINS** (or something in the vicinity) is supported by our ordinary practice of explaining and predicting intentional action via the attribution of attitudes to agents under the assumption of their rationality. If two rational agents’ being of the same mind did not correlate with their intentionally acting alike, then the very idea “that people act the way they do because of what they believe” would seem out of place. To be sure, **TWINS** encapsulates a form of internalism about reasons that is not universally accepted. However, it’s an internalism of the most plausible kind for it explicitly concerns motivating reasons for actions, and so, shouldn’t be incompatible with most forms of externalism or objectivism about normative reasons (Singh 2019, Wang, 2025).

A few caveats about **TWINS** are in order before proceeding. First, it’s a principle of ideal rationality. I will offer no precise definition of an ideal reasoner, but they must at least be perfectly aware of which attitudes they hold as well as being maximally compliant with the norms of epistemic and practical rationality. This vague characterisation suffices to avoid a few distracting objections (e.g. cases involving reasoning mistakes due to cognitive or temporal limitations, failure to keep up with one’s evidence, forgetfulness, behaviour motivated by irrationally-held attitudes, akrasia, weakness of the will, etc). In addition, notice that **TWINS** doesn’t presuppose that it’s possible for distinct agents to have all the same attitudes, after all, doing so would seemingly beg the question against proponents of **Privacy**. It’s then helpful to introduce a label applying to agents who share each other’s attitudes to the maximal possible degree which is compatible with there being attitudes towards private thoughts.

Let agents *x* and *y* be *confidants* if (1) they are mutually acknowledged ideally rational agents, (2) they have engaged in an idealised communicative exchange whereby they have mutually expressed all of their attitudes, and, consequentially, (3) whose post-communication attitudes are the result of updating on everything which they have just learned from each other.¹⁶ In other words, confidants are mutually acknowledged rational peers who pool their information together via communication, and so, end up in attitudinal states which result from factoring in everything they have learned from each other. So, confidants have common knowledge of each other’s attitudes. But not only that: since they take each other as peers in all epistemic and practical rationality-relevant respects, they will be moved towards calibrating their own attitudes in light of what they learn from one another. By doing so, their resulting attitudes will be maximally

¹⁵ My formulation of **TWINS** is derived from Hunter’s *Inter-Personal Reasons*: “There can be no difference in what two people have reason to do without a difference in what they believe or desire” (Hunter, 2017, p. 693; 2022, p. 158; see also Cappelen & Dever 2013, p. 52). It’s also inspired by the law-like generalisation which Ninan refers to as *EXPLANATION* (Ninan, 2016) and as *Thesis (IV)* (Ninan, 2021). However, there are some important differences. First, Ninan’s generalisations quantify over actions instead of reasons. Second, they include a *ceteris paribus* clause (“if all else is equal”) instead of being restricted to ideally rational subjects. Analogous considerations apply to Prosser’s (2025) *Weak Thought-Action Principle*. While **TWINS**’s more restricted scope makes it more artificial, it also makes it more precise, which pays off when it is deployed in the argument laid out in the next section.

¹⁶ The concept of *confidants* is an adaptation of Caie’s (2018) concept of *epistemic confidants*, which he deploys in an independent—but not wholly unrelated—investigation on the limits of objective disagreement between agents who have common knowledge of each other’s doxastic state. One key difference is that, given my focus on rational action, I’m as much concerned about non-doxastic attitudes like desire than about doxastic ones like belief—my *confidants* not only believe the same but desire the same as well.

convergent: not only will they know each other's beliefs and desires, but, to the extent that this is possible, they will generally believe and desire the same as each other too.

There's a great deal of simplification in the above story, some of which I'll address in Section 6.3, but this much is enough for us to draw a few important conclusions. First, the degree to which the third-person attitudes of confidants can differ is minimal. Given their status as ideal reasoners and the fact that they will end up in possession of all the same evidence, it cannot be possible for them to harbour much disagreement about objective matters of fact (if any). Plausibly, this holds as much for their third-person doxastic attitudes as it does for their third-person non-doxastic ones, at least to the extent that we may assume third-person desires (e.g. my desire that candidate A win the election) to be subject to similar rational constraints as their doxastic counterparts.¹⁷ Indeed, if we grant the assumption that all considerations relevant to the adoption of third-person attitudes are of a type which can be unproblematically transmitted via communication among ideal rational peers, then it's safe to say that confidants will just have the same third-person beliefs and desires as each other.

While the shareability of third-person attitudes shouldn't pose too much of a problem, indexical attitudes make matters more complicated. So as not to beg the question against **Privacy**, we must be open to the possibility that confidants will differ in their indexical attitudes even if they don't differ in their third-person ones. Still, proponents of both **Symmetry** and **Privacy** can agree with the following: the indexical attitudes of confidants are *maximally coordinated*, where this roughly means that there's a one-to-one mapping which takes attitudes expressible by one with 'I' to attitudes expressible by the other with 'you' (and vice-versa). If, for example, one believes 'I am F', the other believes 'you are F'; if one has a desire expressible as 'I desire to travel abroad', the other has a desire expressible as 'I desire *you* to travel abroad'. And so on.

Thus, confidants have *maximally coordinated* indexical attitudes, i.e. attitudes which differ only by a systematic permutation of indexically-encoded perspectives ('I' for 'you', 'my' for 'your', etc.).¹⁸ Those sympathetic to **Symmetry** will be advised to hold that *maximally coordinated* indexical attitudes just are the same, after all, these will be attitudes towards the same thoughts. Proponents of **Privacy**, on the other hand, will deny that, claiming that *maximally coordinated* indexical attitudes, as intimately related as they may be, are attitudes towards distinct thoughts. The dispute between **Symmetry** and **Privacy** can thus be reframed as a disagreement about whether *maximally coordinated* indexical attitudes ought to, for the purposes of propositional attitude psychology, be identified or not. Granting that both camps accept that confidants have the same third-person attitudes, the deeper manifestation of their disagreement turns on whether we should take confidants to have all the same rational attitudes (third-personal and indexical) or not. This is the core rationale of the Asymmetry Argument, to which we now turn.

5. The Asymmetry Argument

The Asymmetry Argument can be framed as a three-premise inference to the negation of **Symmetry**:

- (1) Ideally rational agents cannot have all the same attitudes if they have distinct reasons [TWINS].

¹⁷ Unfortunately, the literature on the limits of rational non-doxastic disagreement and interpersonal rational deliberation is scarce, so I might be glossing over some important difficulties. To err on the side of caution, bear the following qualification in mind: even if rationality permits confidants to somehow differ in their third-person attitudes (e.g. maybe it's just as rational to be a compatibilist than an incompatibilist), it's much less plausible that rationality could *require* confidants to so differ. If the only types of disagreement which confidants are permitted to harbour are ones which they have no rational obligation to harbour, then all the premises in the Asymmetry Argument (Section 5) remain correct given some minimal precisifications, some of which I'll address in Section 6.3. I'm grateful to Quentin Ruyant for pointing out some of these issues to me.

¹⁸ Since I exclusively consider confidants who are in each other's presence at a time, there's no obvious reason to think that their other indexical attitudes will be any different. I'll come back to related issues in Section 6.3.

- (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).^{19 20}

Since (1) and (3) have been initially motivated in Section 4, I now comment on (2) before turning to consider objections to each premise in Section 6.

To establish (2), we need only a single case involving confidants who are motivated to act distinctly from each other. Purely general considerations show that this should be an easy task. Suppose, for example, that after completely pooling their information, two agents realise that something terrible will happen unless they immediately change the lightbulb above them, a task which requires one to hold the stairs for the other to climb it. Clearly, there’s a sense in which what they will have to do is different - the lightbulb won’t be changed unless they perform different sub-tasks. So, there’s a sense in which even cooperative agents who maximally agree with each other about what needs to be done and about how that ought to be done might be required to act distinctly.²¹ More generally, it’s a truism that cooperation often requires coordination (of distinct actions), not imitation. Under the assumption that agents have distinct reasons when they’re not motivated to perform the same action, then the above is sufficient to support (2).

Still, it’ll be helpful to have a concrete case in mind. Consider this familiar scenario from Perry (1977, p. 494):

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. (Adapted from Lin 2022, p. 2)

Bear Attack has been put to many uses in the literature. For present purposes, it should be seen as a case involving agents who, though broadly of one mind, act distinctly from each other (and rationally so). However, it’s one thing to say of two agents that they’re broadly in agreement with each other (or that they agree about everything that seems to be of present relevance) and another to say that they’re confidants. If Ann and Bob aren’t confidants, then their third-person attitudes might differ in respects which, though seeming irrelevant at first sight, could possibly explain why they don’t react in the same way upon discovering the same fact (e.g. perhaps, as Lin (2022) suggests, Ann and Bill disagree about what is the best action). If that were the case, then **Bear Attack** would be a completely uninteresting case: it would just illustrate how agents who agree about some things but disagree about others might react distinctly in the same circumstances. I’m not sure how tenable that hypothesis is (more on this in 6.2), but since I don’t have a principled way for demarcating which attitudes are relevant and irrelevant, I’ll simply evade it by stipulating that Ann and Bill are confidants. 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 -

¹⁹ In elaborating the Asymmetry Argument, I’ve been influenced by many prior works. The argument’s roots arguably trace back to Perry (1977; 1979). More recently, related reasoning has been developed by Ninan (2016; 2021), García-Carpintero (2017), Valente (2018), Torre (2018), Torre & Weber (2021, 2022), Lin (2022), and Prosser (2005, 2023, 2025). Other works discussing the rational asymmetries between different indexicals include: Stalnaker (2008), Recanati (2016, 2025), Verdejo (2017, 2019, 2020), Bozickovic (2021), Gray (2022).

²⁰ Does **Privacy** follow from the negation of **Symmetry**? A more comprehensive analysis of the Asymmetry Argument would have to consider the alternative conclusion that first-person thoughts do not have absolute truth-values, and so, should be modelled as e.g. self-ascribed properties or centred contents (Lewis, 1979; also Ninan, 2016, 2021; Shaw 2019; Torre & Weber 2021, 2022). I discuss a related “relativistic” idea in 6.3.

²¹ To be sure, there’s another sense in which what they ought to do is the same, namely, change the lightbulb. More on this in 6.2.

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. With those assumptions in place, we can be sure that Ann and Bill's reactions to the bear attack won't be explainable by their having any conflicting attitudes towards any third-person thought. Seen under this light, **Bear Attack** effectively becomes an example of confidants who have distinct reasons. This concludes my case for (2).

6. *Against Asymmetry*

In the following, I look into objections to each of the Asymmetry Argument's premises.

6.1. *Against (1)*

(1), i.e. **TWINS**, is susceptible to at least two types of criticism. First, one may argue that reasons do not supervene *solely* on attitudes. Second, one may accept the supervenience of reasons on attitudes at the intra-personal level (**CHANGES**) while denying that it must also hold inter-personally (**TWINS**).

Consider the hypothesis that reasons not only supervene on agents' attitudes but also on something else. What could that something else be? A natural candidate is what the agent is capable of doing, that is, which actions they can in principle perform. If so, then agents of the same mind might have distinct reasons due to differing in the actions which are available to each of them (Cappelen & Dever, 2013, pp. 49-56).²² I think this proposal has already been successfully refuted (e.g. Valente, 2018; Torre, 2018; Lin, 2022), but rehearsing those arguments can still be worthwhile, for they point to a general reason to be sceptical of any structurally similar proposal.

For starters, consider confidants *x* and *y* in each other's presence at an arbitrary time, both of whom know that *x* has a reason to perform action α at that time, that *x* is capable of performing α , and that *y* is not so capable (perhaps due to physical limitations, lack of know-how, etc). Given the plausible assumption that no ideally rational subject can have a reason to perform an action which they judge impossible for them to perform, it follows that *y* does not have the same reason as *x*, i.e. a reason to perform α . So, a defender of **Symmetry** could suggest, the fact that *x* and *y* differ in their capacity to perform α could by itself explain why they have distinct reasons compatibly with the assumption that they have all the same attitudes (*contra TWINS*).

To see what is wrong with that hypothesis, imagine that, unbeknownst to both *x* and *y*, whatever impediments were the reason for *y*'s being incapable of performing α had somehow been eliminated from their case. For example, if α were the action of calling the police and only *x* could perform it because only his phone had service, then what we should imagine is that, unbeknownst to *x* and *y*, service had just been restored to *y*'s phone. For another example, if α were the action of clapping one's hands and *y* could not do it due to being handcuffed, then we suppose that *y*'s handcuffs had magically just been unlocked. Would these changes in *y*'s practical capacities entail corresponding changes in *y*'s reasons? Plausibly, they would not, for *y* would still wrongly but rationally believe that she cannot perform α , and we have supposed that an ideally rational agent cannot have a reason to do what she thinks is impossible for her to do. This is enough to motivate a fairly general claim, namely, that our reasons are only sensitive to facts about what we can do intermediately via being sensitive to our beliefs about what we can do. As I hope is clear enough, this is just to concede that agents' reasons do in fact supervene only on their attitudes, exactly as **TWINS** proclaims.

²² Interestingly, the hypothesis under consideration is reminiscent of proposals by authors who notoriously are sceptical of the significance of 'I' and irreducibly indexical thoughts (Cappelen & Dever, 2013; Magidor, 2015). It would be surprising if the shareability of first-person thoughts turned out to depend on an argument frequently associated to so-called *de se* scepticism. This would suggest that the shareability of first-person thoughts is at odds with their having distinctive essentially indexical or first-personal features.

For an objection, one might raise the hypothesis that α is a type of action which can in principle only be performed by x - perhaps because the logical form of that action is something like *the action that x performs α* . I'm not sure if this proposal is as much as coherent in and of itself; in any case, it certainly should not get sympathies from a defender of **Symmetry**. For one thing, it effectively appears to invite a notion of privacy to the realm of actions in a way that dangerously resembles how a proponent of **Privacy** applies that notion to a subset of thoughts. Indeed, both the hypothesis under consideration and **Privacy** share the assumption that a fundamental element of propositional attitude psychology and inferential reasoning is private, disagreeing only about whether this element pertains more closely to the reasoning's premises (thoughts) or to their outcomes (actions). Second, even if we granted the assumption that α is an action performable only by x , the same reasoning as in the previous paragraph could be used to show that y could have a reason to perform α if she wrongly but rationally believed that she is x . Just like agents' reasons don't seem to be sensitive to brute facts about what they can do (as opposed to their attitudes about what they can do), they also don't seem to be sensitive to brute facts about their identity (as opposed to their attitudes about who they are). More generally, for any candidate type of fact on which agents' reasons might additionally supervene, it seems possible to tweak the relevant agents' attitudes about these facts in such a way as to show that, ultimately, their reasons exclusively supervene on their attitudes.

For an alternative, and perhaps more radical, challenge against **TWINS**, one could argue that the supervenience of reasons on attitudes holds only for a single agent, not for multiple ones (Hunter 2017; Hunter 2022, pp. 157-9). In our terminology, the hypothesis is that we can adequately accommodate what Bermúdez calls a basic tenet of propositional attitude psychology by accepting **CHANGES** while rejecting **TWINS**. I believe that the structural analogy between **TWINS** and **CHANGES** is enough to show that this "mixed" strategy is untenable: why would time-slices of a single agent be different from multiple agents with regards to the independence between their reasons and their attitudes? In any case, I'll advance a more general challenge to anyone aiming to save **Symmetry** by means of this strategy. In a nutshell, my point is that acceptance of **CHANGES** can by itself give rise to an argument for a type of thought unshareability that in many ways resembles the workings of the Asymmetry Argument. With one exception, however: the new argument concerns temporally indexical thoughts, not first-personal ones.

To recap, **CHANGES** says that one's reasons cannot change if one's attitudes remain the same. Well, either it's possible for all of an agent's attitudes to remain the same from a time to another or it's not. If it's not, then it's because there are attitudes which by their very nature cannot be retained across time, a conclusion which will invariably motivate the introduction of thoughts that are *temporally private*, i.e. apprehensible only at some times. Presumably, this wouldn't sit well with any defence of **Symmetry**, so, instead suppose that an agent's later self could in principle have retained all of her earlier self's attitudes. A good candidate example would be an idealised agent with flawless memory who does nothing from a moment to another but track the passage of time in a wholly predictable way (think of someone watching their favourite movie for the hundredth time and thinking in succession, e.g. 'the bad guy will soon be killed', 'the bad guy is being killed now', 'the bad guy has just been killed'; Titelbaum 2013, p. 232). Now, notice how the intimate relationship between this agent's time-slices' attitudes resembles that between the attitudes of confidants: intuitively, they're just the same attitudes *modulo* a shift in their indexically-encoded perspectives - e.g. across time, 'soon' gets switched for 'now', across agents, 'you' gets switched for 'I', etc. Finally, notice that a single agent's reasons can change even when her attitudes shift only in this minimal sense. For example, suppose that the movie-watcher has a ritual where she always claps when the bad guy falls dead on the floor. She would then have a reason to clap when she thinks 'the bad guy has fallen dead [now]' but not when she earlier thought 'the bad guy will soon fall dead'. Absent reasons to think that her attitudes have changed in any other way over and above being temporally shifted —e.g. the switch from thinking of the bad guy's death as present instead of future—, the change in her reasons will ultimately require distinguishing between temporally indexical attitudes expressed with distinct indexicals even when coordinated via flawless tracking to the passage of time.

The argument above can be framed in a way that highlights its structural similarity to the Asymmetry Argument. Let **Synchrony** stand to temporally indexical thoughts as **Symmetry** stands to first-person ones (i.e. both principles stating

that the corresponding type of indexical thoughts are shareable across times/time-slices or agents). The Asynchrony Argument then is:

(1*) **CHANGES** [an ideally rational agent's reasons cannot change if their attitudes remain the same].

(2*) An agents' reasons can change even when she does nothing but perfectly keep track of the passage of time in a wholly predictable way.

(3*) If **Synchrony** holds, then the attitudes of an agent who does nothing but perfectly keep track of the passage of time in a wholly predictable way remain all the same.

Therefore, **Synchrony** does not hold (from 1*, 2*, and 3*)

I hope it is easy to see that the Asynchrony Argument supports the (temporal) privacy of temporally indexical thoughts in virtually the same way as the Asymmetry Argument supports the privacy of first-person ones. However, it's hard to see how one could have a good rationale for defending **Symmetry** independently of a more general ambition to vindicate the shareability of thoughts as a whole. If that's right, then there's no promise for a mixed strategy which minimises the damage of rejecting **TWINS** on the basis of accepting **CHANGES**. Either both principles must be taken on board—in which case one ought to challenge the arguments' other premises—or both must be rejected—contrary to a basic tenet of propositional attitude psychology, namely, that reasons supervene on attitudes both for a single agent and for multiple ones.

6.2. *Against (2)*

According to (2), some confidants have distinct reasons. As we've seen, (2) can be supported by general considerations having to do with cooperative action as well as by looking into particular cases like **Bear Attack**. Given the simplicity of these arguments, it strikes me as the most difficult premise to deny. One way to do so is to question our intuitive verdicts about sameness of action and, as a consequence, about sameness of reasons.

No doubt there's a sense in which Ann and Bob do *not* act alike in **Bear Attack**. However, there might also be a plausible sense in which they do so, after all, both act in light of a common goal—e.g. protecting Ann from the bear attack. If asked to explain why they acted as they did, for example, both could give similar answers: 'to protect Ann/her/myself'. So, who is it to say that the concept of action and reasons relevant for propositional attitude psychology is one which, applied to **Bear Attack**, entails that Ann and Bob act distinctly as opposed to alike, or have distinct reasons as opposed to the same ones?

Bermúdez (pp. 48-9, pp. 105-6) makes a few observations that point towards the reasoning outlined above. The context is a critical discussion of Perry (1977), who Bermúdez charges of confusing between intentional actions and the bodily movements via which they are implemented:

Perry is assuming that if our bodies move in comparable ways then we are behaving in the same way [...] But psychological explanations explain actions rather than bodily movements, and actions are (at least partially) individuated by their goal. From this perspective it is not obvious that my rolling up in a ball is the same action as your rolling up in a ball. (Bermúdez, 2017a, p. 48-9)

One of Bermúdez's point which strikes me as clearly correct is that the concept of intentional action relevant to propositional attitude psychology admits of distinct action-types having tokens which look just like one another, as plausibly the case when the relevant tokens have radically distinct goals (e.g. an agent walking home might look just like another performing a contemporary dance piece). However, Bermúdez's contention that actions ought (at least partially) to be individuated by their goal suggests more than that, namely, that action-tokens which are radically unlike can be

grouped under the same type when they have the same goal. This more general point is echoed in further remarks by Bermúdez elsewhere in his book:

What I am doing is saving myself from the bear. This action-type can have other tokens, such as, for example, my shooting the bear. (p. 57)

The real issue [...] is not who is performing the action, but what the goal of the action is. What matters for the thought that *I myself am being pursued by a bear* is that it should give rise to actions with the aim of extricating *me* from my predicament. Those might be actions that I perform —standing my ground and making noise (recommended), or running away (not recommended). Or they might be actions that someone else could perform —distracting the bear or attacking it with bear spray. (Bermúdez, 2017a, p. 105)

On one way of reading the above, Bermúdez's suggestion is that, for the purposes of propositional attitude psychology, we neither need to differentiate the actions performed by Ann and Bob nor their underlying motivating reasons. Regardless of the oddity of saying that Ann's curling up into a ball and Bob's running for help are tokens of the same action-type, I don't think this is an absurd claim. Still, absurd or not, it cannot suffice to motivate the rejection of (2).

A general hypothesis suggested by Bermúdez's remarks is that confidants need not be attributed distinct reasons because, though there might be superficial differences in how they act, their acting in light of the same goal entails that their actions (and reasons) can appropriately be grouped under the same type. An initial question is: how should goals be individuated (and why should that be easier than individuating actions/reasons)? Surely, if one is creative enough, one can readily find a general enough action-type under which to group any two arbitrary action-tokens that have little to do with each other. And surely, in the absence of a more precise theory of goal individuation, one's creativity will allow one to readily do the same for agents' goals as well. One then wonders whether shifting from action/reason-talk to goals-talk will be much of an improvement.

Another question is whether propositional attitude psychology sits well with the grouping together of action-tokens whenever they have the same goal. One obvious issue with that idea is that it would restrict the scope of psychological generalisations in a dramatic way, entailing that whatever differences there are between, for example, Ann and Bob's reactions to the approaching bear, cannot be given an explanation in attitudinal terms. In other words, Ann and Bob's attitudes would at most allow us to explain why their goal is to protect Ann whilst being completely silent about why Ann ought to do so by curling up into a ball and Bob, by running for help. This conflicts with the general idea that, in practical reasoning, an agent's attitudes stand to their actions just like the premises of a theoretical inference stand to its conclusion.²³ To spin the same point in yet another way: if the practical implications of being in Ann's attitudinal state were the same as that of being in Bob's, then their attitudinal states would underdetermine whether they ought to curl up into a ball or to run for help. How, then, could Ann rationally explain why she curled up into a ball as opposed to running for help if even a full disclosure of her attitudinal state wouldn't justify anything more specific than that she ought act with the goal of protecting Ann from the bear?

To be sure, all the above is compatible with Bermúdez's fair remark that any sufficiently interesting concept of intentional action must be compatible with there being permissible variations on the manner in which it can be performed. Perhaps we wouldn't need to say that Ann acted distinctly if, instead of curling up into a ball head first, she started by curling up her legs. However, this weaker claim is surely not enough to entail anything as strong as the type-identity between Ann and Bob's actions. I'm also tempted to agree with Bermúdez that no two action-tokens, however similar to each other, ought to be grouped under the same type if they uncontroversially have distinct goals (at least not when our primary concern is attitudinal explanations of action; cf. Evans, 1982, p. 203-4). But this is compatible with there sometimes being a need to distinguish between action-tokens which, intuitively, have the same goal, and so, to attribute dis-

²³ As I read him, Bermúdez (p. 47) explicitly endorses this general idea and goes as far as deploying it in an argument against Perry.

tinct reasons to agents even when their actions are perfectly orchestrated as part of a joint plan. If so, then (2) remains in good-standing.

6.3. *Against (3)*

Since (3) involves a particular interpretation of what the shareability of first-person thoughts entails, it looks like the most promising target for proponents of **Symmetry**. To recap, (3) holds that, if **Symmetry** is true, then confidants have all the same attitudes. I take (3) to follow from the following four premises:

- (ii) Two agents have all the same attitudes if and only if they apprehend all the same thoughts and adopt the same attitude-types towards them
- (iii) If **Symmetry** is true, then confidants apprehend all the same *first-person* thoughts and adopt the same attitude-types towards them
- (iv) Confidants apprehend all the same *third-person* thoughts and adopt the same attitude-types towards them
- (v) Attitudes are either third-person or first-person

Therefore, if **Symmetry** is true, then confidants have all the same attitudes.

Given the above, a defender of **Symmetry** is advised to challenge at least one of the four premises supporting (3). For ease of exposition, let's focus on those two attitudes of Ann and Bob which most see as the most relevant in **Bear Attack**:

(AA): Ann's attitude expressible as 'I am about to be attacked by a bear'

(AB): Bob's attitude expressible as 'you are about to be attacked by a bear'²⁴

(i) can be challenged by arguing that there's more to sharing an attitude than adopting the same attitude-type towards the same thought. For example, attitudes like belief might be triples involving the attitude-type of belief, a thought, and some third-ingredient under which the thought is believed (e.g. a mode of presentation, perspective, guise, etc).²⁵ If so, then AA and AB might be equal with respect to the first two elements while differing in their third one. Alternatively, one might accept that attitudes are pairs of attitude-types and thoughts but deny (ii) by holding that coarse-grained attitude-types such as belief and desire are too simplistic, and so, must give way to fine-grained attitude-types which correspond to different ways of believing and desiring the same thought.²⁶ For example, AA and AB might differ because the former involves a first-person type of belief but not the latter. Clearly, both strategies are structurally analogous in an important sense: they attempt to save **Symmetry** by holding that AA and AB are different attitudes regardless of involving the same thought. For this very reason, both strategies suffer from the same problem with respect to their compatibility with **Symmetry**.

Suppose AA and AB are different attitudes due to the denial of either (i) or (ii). Faced with this objection, a proponent of **Privacy** might ask: but are attitudes shareable or private on this new view under consideration? Surely, a defender of **Symmetry** must hold that they're shareable, for any view where some attitudes can in principle only be held by some agents is anathema to the spirit of thought shareability. But if they are, then it must in principle be possible for Bob to have AA in **Bear Attack** (and Ann to have AB). By assumption, Bob doesn't have AA. Furthermore, he's maximally ra-

²⁴ Though AA and AB are doxastic attitudes, my reasoning below could have equally been framed in terms of non-doxastic indexical attitudes like desire.

²⁵ The view under consideration is reminiscent of Perry's (1977) and has been recently developed by Prosser (2025). Since Perry is one of Bermúdez's (p. 46-50) main targets of criticism, it's fair to assume that he's not the least sympathetic to it.

²⁶ A similar idea is elaborated for temporally indexical attitudes by Shaw (2019).

tional. So, the reason why Bob doesn't have AA is because it would have been irrational for him to do so. If that's the case, then rationality appears to require Ann to have AA while forbidding Bob to do so. Given that they're confidants, this can only be because the rationality of holding an attitude towards a thought is an agent-relative matter which can vary across agents even when all else remains the same. In other words, the proposals are committed to:

Relativism: though confidants could in principle have all the same attitudes towards the same thoughts, it could be irrational for them to do so.

Whatever the merits of this principle, I'm confident that **Relativism** undercuts the fundamental motivation for anyone aiming to defend **Symmetry** in the first place, and so, has no place in a view which emphasises the shareability of thoughts. To see why, suppose that **Relativism** is true, and so, that Ann and Bob apprehend the same indexical thoughts as each other regardless of being required to do it in distinct ways. Now, consider: why would the shareability of thoughts allowed by that type of view amount to a vindication of **Symmetry** over **Privacy** as opposed to a pyrrhic victory?

My worry is that a proponent of **Privacy** would be justified in taking the conjunction of **Symmetry** and **Relativism** (henceforth, **Symmetry-Relativism**) as, at best, a close cousin to their own view, at worst, a notational variant thereof. For one thing, defenders of **Privacy** and of **Symmetry-Relativism** agree that Ann and Bob cannot rationally share AA and AB —be this because AA and AB involve private thoughts or because they involve thoughts which, though shareable, cannot rationally be apprehended in the same way by anyone. If they disagree at all, they do so only with respect to whether AA and AB involve the same thought. But it's not clear whether this purported disagreement amounts to anything substantial. Having rejected the standard view according to which attitudes are pairs of coarse-grained attitude-types and thoughts, the defender of **Symmetry-Relativism** doesn't appear to mean anything by, 'AA and AB involve the same thought', over and above the claim that Ann's holding AA and Bob's holding AB involves them apprehending thoughts which could exemplify a successful instance of knowledge transmission via testimony. However, as we've seen in Section 3, this is a claim that a proponent of **Privacy** will be more than happy to embrace. A defender of **Symmetry-Relativism** might retort that her view fares better due to eschewing any hint of privacy in propositional attitude psychology. But her opponent could simply point out that **Relativism** does introduce a type of privacy, after all, it attributes a sort of agent-relativity to the rationality of holding an attitude which obviously resembles the asymmetry in the conditions of thoughts' apprehension countenanced by **Privacy**. Against this, proponents of **Symmetry-Relativism** might insist that they're the only ones able to maintain that attitudes expressed with 'you' and 'I' like AA and AB might literally be said to be attitudes towards *the same thought*. However, they can do so only by accepting that this true shareable thought is not one which Ann and Bob can be said to believe in the same way. Surely, the proponent of **Privacy** will exclaim, "ways of believing a thought" are neither less nor more mysterious theoretical posits than their favoured private thoughts. Furthermore, **Symmetry-Relativism** is committed to there being attitudes which, though accessible to anyone in principle, are not *rationally* accessible to everyone. Indeed, the proponent of **Symmetry-Relativism**'s claim that Ann and Bob could in principle share AA and AB —though not rationally— reminds me of W. W. Jacobs' horror story *The Monkey's Paw*, where a man who asks a demon for £200 is granted his wish only through receiving the insurance compensation for the sudden death of his loved one. Shareable, yes, but at what cost?

Finally, one who defends **Symmetry-Relativism** will still be hard-pressed to explain how the rationality of holding an attitude could be agent-relative in that sense. If AA is a doxastic attitude towards a true thought which Ann would be irrational not to hold, then, presumably, Ann's holding AA amounts to her being in possession of a piece of knowledge. But since Bob would be irrational for holding AA, then Bob's having AA could not amount to his being in possession of a piece of knowledge - for how could rationality forbid Bob from having knowledge? This is indication that **Symmetry-Relativism** will ultimately turn out to be committed to a particularly nasty form of epistemic relativism wherein some pieces of knowledge are in principle accessible only to particular agents. How can any similar conclusion be an acceptable corollary of **Symmetry**, a view predicated on the need to vindicate the objectivity and shareability of thoughts across multiple perspectives?

In summary, by rejecting (i) or (ii), the proponent of **Symmetry** will have to buy onto **Relativism**. As a consequence, they will ultimately have to accept that propositional attitude psychology and knowledge are infused with a degree of agent-relativity that at the very least is as worrisome as any implications of postulating private thoughts. I take all this to suggest that **Symmetry-Relativism** and **Privacy**'s purported disagreement on the shareability of first-person thoughts is less substantial than it might initially appear. If so, then the denial of (i) and (ii) is less a way to save **Symmetry** than of conceding that distinct agents cannot in principle adopt the same attitudes towards the same thoughts, as proponents of **Privacy** have been defending all along, though not in these exact terms.

Let us now consider (iii) and (iv). To deny (iii), one must hold that confidants can differ in their third-person attitudes. On a weaker reading, the claim is that confidants are rationally permitted to hold conflicting attitudes towards some of the same third-person thoughts. On the stronger reading, they not only have permission to disagree but are required to do so. I take the stronger reading to be strongly implausible. Even those with highly permissive views of rationality are careful in saying that, due to evidential underdetermination or related phenomena, a single body of evidence might *permit* more than a single attitude towards the same thought. However, if confidants are permitted (but not required to) harbour third-person disagreements, then they are equally permitted to agree about these same matters. Thus, we need only stipulate that, through rational deliberation, confidants like Ann and Bob have decided to adopt a fully conciliatory approach to any matters about which rationality is permissive, and so, have preempted the possibility of their ever making different choices in cases where rationality allowed them multiple options. In any case, it's just immediately implausible that Ann and Bob's distinct actions in **Bear Attack** are the product of their holding permissibly distinct attitudes towards a single third-person thought; the fact that they ought to act differently from each other is plausibly not a product of a decision which they could rationally have not taken —it's just an inevitable consequence of their need to coordinate their actions in light of their shared goal. So, there's no reason to think that permissible disagreements might underlie Ann and Bob's distinct actions.

Finally, one might deny (iv) by holding that some attitudes are neither third-person nor first-person. This would allow one to hold that Ann and Bob don't have the same attitudes regardless of having the same third-person and first-person ones. But which other attitudes could those be? One hypothesis is that they differ in their demonstrative attitudes, those naturally expressible using 'this' or 'that'. However, we may simply stipulate that Ann and Bob are acquainted with all the same objects, and so, able to demonstratively pick out the same things as one another. Of course, a defender of **Symmetry** won't find solace in that view once defended by Bertrand Russell whereby we can only be directly acquainted with ourselves and the inner objects of our own consciousness. This view would surely entail that Ann and Bob don't have the same demonstrative attitudes, but only in detriment to the shareability of demonstrative thoughts.

Alternatively, one might raise the hypothesis that the relevant difference between Ann and Bob lies in their being in distinct perceptual states. For example, Ann might be seeing the approaching bear's face while Bob can only see its back. Surely, this cannot be the whole explanation of why they act differently from each other. In any case, we might circumvent this complication by tweaking **Bear Attack** in such a way as to minimise any perceptual discrepancies between the agents (e.g. picture Ann and Bob in a sensory deprivation chamber, being telepathically informed that something horrible will happen unless Ann mentally counts from 1 to 20, and Bob mentally counts from 20 to 1). Additionally, one might take their relevant difference to be a matter of their emotions or affective states, after all, isn't Ann bound to be much more scared of the approaching bear than Bob is? That might be so, but cannot we easily come up with a boring case involving confidants who ought to act differently due only to their sense of duty, and so, who comply with their reasons in the utmost blasé manner?²⁷ There surely are other hypotheses I could consider —e.g. their values, their preferences, their habits, their tastes— but I take the above to sufficiently justify the conclusion which from the outset struck me as the most intuitive one, namely, that Ann and Bob act differently from each other because of asymmetries in their attitudes expressible with 'you' and 'I', those very asymmetries which a proponent of **Privacy** is well-positioned to accommodate by taking some of them as targeting private thoughts which can only be apprehended by their very owners.

²⁷ The irrelevance of perceptual and emotional discrepancies in **Bear Attack** has been, correctly in my view, pointed out by Gray (2022).

7. Conclusion

I have claimed that the shareability of first-person thoughts doesn’t sit well with the rational asymmetries between ‘you’ and ‘I’ showcased by the Asymmetry Argument. Given that confidants, agents whose attitudes agree with one another’s to the maximal degree, can have distinct motivating reasons for action, it seems that we have no better way to explain how that is possible without accepting that thoughts and attitudes are subject to some degree of privacy and unshareability. In that context, **Privacy** emerges as a natural candidate to account for the relevant data: since first-person thoughts are private, then even confidants will differ in which first-person attitudes they hold. I looked into multiple ways for trying to make **Symmetry** compatible with these findings, but none turned out to be ultimately tenable. Some saddled **Symmetry** with the rejection of a basic tenet of propositional attitude psychology (6.1), some with an implausible picture of how actions and reasons ought to be individuated (6.2), and others with an unstable commitment to **Relativism** which appears anathema to the idea that ‘you’ and ‘I’ can underly the holding of the same attitude (6.3). I might have failed to identify a further sense in which the shareability of first-person thoughts would both be philosophically significant and incompatible with **Privacy**, but I expect to have said enough to show that the burden of proof lies on defenders of **Symmetry**’s side. For the time being, my conclusion is that either we must accept that first-person thoughts are private or accept that they are only shareable in a weak sense which any defender of **Privacy** could be happy to embrace. If I’m right, then the thesis that first-person thoughts are shareable is either false or uninteresting.²⁸

Acknowledgements

I would like to express my sincere gratitude to Víctor M. Verdejo, Valen Simpson, Quentin Ruyant, and two anonymous referees for invaluable feedback on this paper. I’m also thankful to audiences at the Valencia Philosophy Lab, LOGOS, Indexical Dynamics colloquium (Collège de France), Language & Mind group (Arché, St Andrews), University of Aberdeen, and 6th PLM Workshop (University of Vienna). The research leading to this paper has been funded by: FCT Researcher Fellowship (University of Lisbon, Center of Philosophy, LanCog), Margarita Salas fellowship (Universitat de Barcelona), and the projects Philosophy of Hybrid Representations (PID2020-119588GB-I00) and Eliminativism, Fictionalism and Expressivism (PID2019-106420GA-I00).

REFERENCES

- Alvarez, M. & Way, J. (2024). Reasons for action: Justification, motivation, explanation. *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <<https://plato.stanford.edu/archives/fall2024/entries/reasons-just-vs-expl/>>.
- Bermúdez, J. L. (2005). Evans and the sense of “I”. In J. L. Bermúdez (Ed.). *Thought, reference, and experience: Themes from the philosophy of Gareth Evans* (pp. 164-94). Oxford University Press.

²⁸ One interesting avenue of research I haven’t explored concerns the implications of Bermúdez’s commitment to **Essential Indexicality** (see footnote 4) and **Symmetry**. *Prima facie*, **Essential Indexicality** draws a sharp divide between attitudes expressed with ‘I’ from all others, including those which, according to **Symmetry**, are attitudes to the same thought. The dissonance between **Essential Indexicality** and **Symmetry** might mean one of two things: either it shows the existence of a deep internal conflict in Bermúdez’s account, or it hides the key to dissolving the arguments I presented in this paper. I wouldn’t be surprised if the second option turns out to be the correct one.

- Bermúdez, J.L. (2011). Self-knowledge and the sense of “I”. In A. Hatzimoysis (Ed.). *Self-knowledge* (pp. 226-45). Oxford University Press.
- Bermúdez, J. L. (2017a). *Understanding “I”: Language and thought*. Oxford University Press.
- Bermúdez, J. L. (2017b). Yes, essential indexicals really are essential. *Analysis*, 77(4), 690-694.
- Bermúdez, J. L. (2019). First person thoughts: Shareability and symmetry. *Grazer Philosophische Studien*, 96(4), 629-638.
- Bozickovic, V. (2021). *The indexical point of view: On cognitive significance and cognitive dynamics*. Routledge.
- Byrne, A. & Thau, M. (1996). In defence of the Hybrid View. *Mind*, 105(417), 139-149.
- Cappelen, H. & Dever, J. (2013). *The inessential indexical: On the philosophical insignificance of perspective and the first person*. Oxford University Press.
- Caie, M. (2018). Agreement and updating for self-locating belief. *Journal of Philosophical Logic*, 47(3), 513-547.
- Cumming, S. (2013). From coordination to content. *Philosophers’ Imprint*, 13(4).
- Dummett, M. (1980). *The interpretation of Frege’s philosophy*. Harvard University Press.
- Evans, G. (1981). Understanding demonstratives. In H. Parret (Ed.). *Meaning and understanding* (pp. 280–304). Clarendon Press.
- Evans, G. (1982). *The varieties of reference*. Oxford University Press.
- Fine, K. (2007). *Semantic relationism*. Blackwell.
- Frege, G. (1956). The thought: A logical inquiry. *Mind*, 65(259), 289-311.
- García-Carpintero, M. (2017). The philosophical significance of the de se. *Inquiry*, 60(3), 253-276.
- Goodman, R. (2025). Shared thought and communication. In J. Bermúdez, M. Valente & V. M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 218-239). Oxford University Press.
- Gray, A. (2022). Minimal fregeanism. *Mind*, 131(522), 429-458.
- Gray, A. (2025). Thinking the Same-ish. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 261-286). Oxford: Oxford University Press.
- Heck, R. K. (2002). Do demonstratives have senses?. *Philosophers’ Imprint*, 2(2), 1-33.
- Hunter, D. (2017). Practical reasoning and the first person. *Philosophia*, 45(2), 677-700.
- Hunter, D. (2022). *On believing: Being right in a world of possibilities*. Oxford University Press.
- Lewis, D. (1979). Attitudes de dicto and de se. *Philosophical Review*, 88(4), 513-543.
- Lin, L. (2022). Attitudes and action: Against de se exceptionalism. *Inquiry*. Advance online publication: <https://doi.org/10.1080/0020174X.2022.2158126>
- Loar, B. (1976). The semantics of singular terms. *Philosophical Studies*, 30(6), 353-377.
- Longworth, G. (2013). IV—Sharing thoughts about oneself. *Proceedings of the Aristotelian Society*, 113(1pt1), 57-81.
- Longworth, G. (2014). You and me. *Philosophical Explorations*, 17(3), 289-303.
- Magidor, O. (2015). The myth of the de se. *Philosophical Perspectives*, 29(1), 249-283.

- McDowell, J. (1984). De re senses. *Philosophical Quarterly*, 34(136), 283-294.
- Morgan, D. (2009). Can you think my 'I' thoughts?. *Philosophical Quarterly*, 59(234), 68-85.
- Morgan, D. (2024). Islands of perspectival thought: A case study. *Ergo: An Open Access Journal of Philosophy*, 10: 44.
- Ninan, D. (2016). What is the problem of de se attitudes?. In S. Torre & M. García-Carpintero (Eds.). *About oneself: De se thought and communication* (pp. 86-120). Oxford University Press.
- Ninan, D. (2021). De se attitudes and action. In H. Geirsson & S. Biggs (Eds.). *The Routledge handbook of linguistic reference* (pp. 482-498). Routledge.
- Onofri, A. (2019). Loar's puzzle, similarity, and knowledge of reference. *Manuscrito*, 42(2), 1-45.
- Perry, J. (1977). Frege on demonstratives. *Philosophical Review*, 86(4), 474-497.
- Perry, J. (1979). The problem of the essential indexical. *Noûs*, 13, 3-21.
- Prosser, S. (2005). Cognitive dynamics and indexicals. *Mind and Language*, 20(4), 369-391.
- Prosser, S. (2023). Tense and emotion. In K. M. Jaszczolt (Ed.). *Understanding human time* (pp. 11-29). Oxford University Press.
- Prosser, S. (2025). Shared egocentric thoughts. In J. Bermúdez, M. Valente & V. M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 52-74). Oxford University Press.
- Recanati, F. (2016). *Mental files in flux*. Oxford University Press.
- Recanati, F. (2025). Sameness of mode of presentation. In J. Bermúdez, M. Valente and V.M. Verdejo (Eds.). *Sharing thoughts: Philosophical perspectives on communication and intersubjectivity* (pp. 193-217). Oxford University Press.
- Shaw, J. R. (2019). De se exceptionalism and frege puzzles. *Ergo: An Open Access Journal of Philosophy*, 6, 1057-1086.
- Singh, K. (2019). Acting and believing under the guise of normative reasons. *Philosophy and Phenomenological Research*, 99(2), 409-430.
- Stanley, J. (2011). *Know how*. Oxford University Press.
- Stalnaker, R. (2008). *Our knowledge of the internal world*. Oxford University Press.
- Titelbaum, M. G. (2013). *Quitting certainties*. Oxford University Press.
- Torre, S. (2018). In defense of de se content. *Philosophy and Phenomenological Research*, 97(1), 172-189.
- Torre, S. & Weber, C. (2021). What is special about de se attitudes?. In H. Geirsson & S. Biggs (Eds.). *The Routledge handbook of linguistic reference* (pp. 464-481). Routledge.
- Torre, S. & Weber, C. (2022). De se puzzles and frege puzzles. *Inquiry*, 65(1), 50-76.
- Valente, M. (2018). What is special about indexical attitudes?. *Inquiry*, 61(7), 692-712.
- Valente, M. (2021). On successful communication, intentions and false beliefs. *Theoria*, 87(1), 167-186.
- Valente, M. & Onofri, A. (2023). A puzzle about communication. *Review of Philosophy and Psychology*, 14(3), 1035-1054.
- Valente, M. & Verdejo, V. M. (2021). Relationism and the Problem of Publicity. *Pacific Philosophical Quarterly* 103(3), 645-669.
- Verdejo, V. M. (2017). De se content and action generalisation. *Philosophical Papers*, 46(2), 315-344.

- Verdejo, V. M. (2019). The second person perspective. *Erkenntnis*, 86(6), 1693-1711.
- Verdejo, V. M. (2020). Rip Van Winkle and the retention of 'today'-belief: A puzzle. *Res Philosophica*, 97(3), 459-469.
- 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>
- Wang, L. (2025). Taking motivating reasons' deliberative role seriously. *Philosophical Studies*, 182, 725-744.
- Weber, C. (2014). Indexical beliefs and communication: Against Stalnaker on self-location. *Philosophy and Phenomenological Research*, 90(3), 640-663.

MATHEUS VALENTE is an Assistant Professor at UNED, Spain, who specializes in the philosophy of mind, language, and epistemology. He has published about indexical and self-locating attitudes, communication, reference, and demonstrative thoughts.

ADDRESS: UNED. Humanidades. Paseo de Senda del rey, 7 (28040 Madrid, Spain). Email: matheusvalenteite@gmail.com – ORCID: 0000-0001-6380-2623



FROM IMMEDIATE PERCEPTION TO PERCEPTUAL BELIEF

(*De la percepción inmediata a la creencia perceptiva*)

Derek H. Brown*

University of Glasgow

<https://orcid.org/0000-0002-5765-2443>

Keywords

Direct perception
Indirect perception
Immediate perception
Mediate perception
Demonstratives
Perceptual demonstratives
Deferred demonstration
Ostension
Perceptual belief
Sense-datum theory
Naturalized sense-datum theory

ABSTRACT: Pretheoretically, our senses give us immediate access to the world around us. Perceptual demonstratives (e.g. uttering “that is a bottle” while pointing at a bottle) seem to directly refer to their intended referents. Should we explain such referential directness by perceptual immediacy? Bermúdez (2000) argues that we should, and offers a powerful theory of immediate and mediate perception (Naturalized sense-datum theory) to explain how. He also argues that a different but influential theory of immediate and mediate perception (that found in traditional sense-datum theory) cannot explain referential directness by perceptual immediacy. I argue that his criticisms fail in instructive ways. Following this I assess the extent to which Bermúdez’s theory can explain a critical perceptual challenge, that of explaining the experiential similarities between perceptual experiences with varying degrees of veridicality. I argue that Bermúdez’s theory can only meet this challenge by embracing a veil of perception akin to that of his opponent – something he sought to avoid. Before concluding, I present the underlying framework of this debate devoid from any appeal to sense-datum theories. The reason is because that framework has considerable power and relevance to a number of contemporary issues in philosophy of perception.

Palabras clave

Percepción directa
Percepción indirecta
Percepción inmediata
Percepción mediata
Demostrativos
Demostrativos perceptuales
Demostración diferida
Ostensión
Creencia perceptiva
Teoría de datos sensoriales
Teoría naturalizada de datos sensoriales

RESUMEN: Preteóricamente, nuestros sentidos nos proporcionan acceso inmediato al mundo que nos rodea. Los demostrativos perceptuales (e.g. proferir «Eso es una botella» mientras se señala una botella) parecen referir directamente a sus referentes. ¿Deberíamos explicar este carácter directo de su referencia por su inmediatez perceptiva? Bermúdez (2000) defiende que sí, y ofrece una poderosa teoría de la percepción inmediata y mediata para explicar cómo (la teoría de los datos sensoriales naturalizados). También sostiene que otra teoría distinta, pero influyente, de la percepción inmediata y mediata (la que se encuentra en la teoría tradicional de datos sensoriales) no puede explicar el carácter directo de la referencia por su inmediatez perceptiva. En este trabajo, argumento que sus críticas fallan de manera instructiva. A continuación, evalúo en qué medida la teoría de Bermúdez puede explicar un desafío perceptual crítico: explicar las semejanzas experienciales entre experiencias perceptivas con diferentes grados de veracidad. Argumento que la teoría de Bermúdez solo puede abordar este desafío adoptando un velo de la percepción similar al de su oponente – algo que intentaba evitar. Antes de concluir, presento el marco subyacente a este debate, desprovisto de menciones a las teorías de datos sensoriales. La razón para esto es porque este marco tiene un poder considerable y resulta relevante para una serie de cuestiones contemporáneas en filosofía de la percepción.

* **Correspondence to:** Derek H. Brown. School of Humanities, University of Glasgow, G12 8QQ, UK. – derek.brown.2@glasgow.ac.uk – <https://orcid.org/0000-0002-5765-2443>

How to cite: Brown, Derek 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>).

Received: 14 February, 2024; Final version: 17 June, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Gako-hitzak

Zuzeneko pertzepzioa
 Zeharkako pertzepzioa
 Berehalako pertzepzioa
 Pertzepzio bitartekotua
 Erakusleak
 Pertzepziozko erakusleak
 Frogantza geroratua
 Ostentsioa
 Pertzepziozko ustea
 Zentzumen-datuen teoria
 Zentzumen-datuen teoria naturalizatua

LABURPENA: Ikuspegi teoriko ororen aurretik, gure zentzumenek zuzenean jartzen gaituzte inguratzen gaituen munduarekin harremanetan. Badirudi pertzepziozko erakusleak (adibidez, «Hori botila bat da» esatea botila bat seinalatuz) zuzenean erreferentzia egiten ari zaizkiela adierazi nahi diren erreferenteci. Azaldu behar al genuke zuzentasun erreferentzial hori pertzepziozko berehalakotasunaren bitartez? Bermúdezek (2000) baietz dio, eta, nola azaltzeko, pertzepzio berehalako eta bitartekotuaren teoria senda bat eskaintzen du: zentzumen-datuen teoria naturalizatua. Halaber dio pertzepzio berehalako eta bitartekotuaren beste teoria eragin handiko batek (zentzumen-datuen teoria tradizionalakoak) ezin duela azaldu zuzentasun erreferentziala pertzepziozko berehalakotasunaren bitartez. Argudiatzen dut haren kritikek modu esanguratsuan huts egiten dutela. Ondoren, ebaluatuko dut zer neurritan azaldu dezakeen Bermúdezen teoriak pertzepziozko erronka erabakigarri bat: egiazkotasun-maila desberdinetako pertzepziozko esperientzien artean dauden esperientziatzko antzekotasunak azaltzeari dagokiona. Argudiatuko dut Bermúdezen teoriak aurkariarenaren pareko pertzepziozko belo bat onartuz bakarrik gaindi dezakeela erronka hori, bera hori saihesten ahalegindu zen arren. Amaitu aurretik, eztabaida honen azpiko markoa aurkeztuko dut, zentzumen-datuen teorien aipurik gabe. Arrazoi da marko horrek indar eta garrantzi handia duela gaur egungo pertzepzioaren filosofiako zenbait auzitarako.

How direct or immediate is our access to the physical world? It seems undeniable that perception is our most basic epistemic and informational route to the physical world. Thoughts and speech can be about the physical world, but when they are, their accuracy relies on what we see, hear, and touch in the world. Perception is our most basic route. How direct or immediate is our perceptual access to the physical world? The twentieth century was largely dominated by a view, sense-datum theory, according to which our perceptual access to the world is mediated by the perception of non-physical mind-dependent objects. Many today find this unpalatable. This also seems to conflict with a kind of directness in thought and speech that we arguably can have about physical objects. When I look at a bottle and think or say ‘That is a bottle’, ‘that’ seems to directly refer to that object. How can such referential directness be explained, if perception of the physical world is mediated? We can turn this question on its head: ideally, shouldn’t we explain such referential directness by our ability to immediately perceive the physical world, particularly given that perception affords our most basic access to that world? If so, what might such an account look like?

That is the topic of Bermúdez (2000).¹ His proposed solution posits a tight explanatory connection between perceptual immediacy and direct demonstrative thought or belief. However, he argues, while we cannot immediately perceive physical objects, we can immediately perceive their front-facing surfaces. He calls the view *Naturalized sense datum theory* (NSD). My interest in this topic is twofold. First, I think the issues just outline are foundational, and have not received enough attention in recent years. I thus wish to demonstrate their contemporary interest. Second, while there are many merits to Bermúdez’s account, I think his argument fails, and that NSD is ultimately forced to accept a kind of veil of perception that he sought to avoid. Thus, overall, I defend a veil of perception, but in a way that explains how we can directly refer to physical objects via perceptual demonstratives.

I proceed as follows. I first describe and defend the proposed connection between immediate perception and direct demonstrative reference (section 1) and then outline Bermúdez’s overall argument for NSD (section 2). Following this, I explain and motivate Bermúdez’s NSD (section 3), and why he thinks traditional sense datum theory (SDT) cannot explain direct demonstrative reference (section 4). I then respond to his critique of SDT (section 5). That completes my critique of Bermúdez’s overall argument, the upshot of which is that there are different, viable conceptions of immediate perception that can be used to explain how we directly demonstrate physical objects. I take this to be of value on its own. In the rest of this paper, I examine the extent to which NSD can address The Problem of Perception. The reason is because a balanced debate between NSD and SDT theorists requires consideration, not

¹ All references to Bermúdez are to his (2000) unless otherwise specified.

only of a NSD-friendly phenomenon like perceptual demonstratives, but also of a SDT-friendly one like The Problem of Perception.

The Problem of Perception is the challenge of explaining the differences between veridical, illusory, and hallucinatory perceptions while respecting the fact that these disparate types of perceptions can be subjectively indistinguishable. This is a foundational perceptual problem in its own right. It is particularly important in a discussion about perceptual immediacy and different versions of sense datum theory, because sense datum theory is (rightly) regarded as offering a powerful, though controversial, solution to this problem. I argue, however, that NSD cannot solve The Problem of Perception, and that The Problem creates an internal tension for NSD that is difficult to resolve (section 6). I then consider a resolution by adding nonconceptual content to NSD, nonconceptual content being something that Bermúdez (alongside others) has done much to articulate and defend (section 7). Bringing nonconceptual content to bear on these issues is also of interest because nonconceptual content has many current advocates. I argue that this addition can indeed solve The Problem of Perception, but at the cost of reverting to a kind of perceptual veil that NSD sought to avoid. The upshot of this part of the paper is therefore that NSD cannot, in the end, avoid a veil of perception. I briefly connect this discussion to the recent interest in naïve realism, which offers yet another conception of immediate perception (section 8). In the penultimate section (section 9), I lay bare the underlying framework of this discussion, divorced from the constraints of sense-datum theories (SDT and NSD). I indicate how this framework has numerous unexplored applications, and then offer a brief conclusion.

What I hope to convince readers of is the importance of examining perceptual immediacy and perceptual mediation, both on their own and in relation to directness in thought and speech. Bermúdez's Naturalized sense datum theory provides a fascinating means of facilitating this discussion. Following Bermúdez, I focus on visual perception throughout.

1. *The Reference Constraint*

The notion of “indirect” or “mediated” perception has a long, distinguished history, with numerous philosophers arguing that we “directly” or “immediately” perceive ideas or sensations and “indirectly” or “mediately” perceive the surrounding physical world. Following Bermúdez (see also e.g. Jackson, 1977; Snowdon, 1992), I distinguish between an epistemic and a perceptual version of this distinction. To start:

Perceptual distinction: When one perceives something *O* by virtue of perceiving something else, one mediately perceives *O*. When a perception of a thing isn't mediated, it is immediate.

Epistemological distinction: Beliefs about perceptions that are fully justified by those perceptions are direct perceptual beliefs, and those that are not indirect perceptual beliefs.²

The perceptual distinction applies to perceptual states on their own, and the epistemological distinction is centred on the kinds of justification that perceptual states can provide for perceptual beliefs. The former is about the structure or metaphysics of perceptual states themselves, and the latter is about how perceptual states feed into associated epistemic states.³ The kinds of perceptual states I focus on are perceptual experiences (or instances of perceptual consciousness), as I take these to be what philosophers of perception are most often concerned with. Let me first elaborate on the epistemological distinction.

² One could formulate this in terms of knowledge instead of beliefs if one preferred.

³ Lyons (2023) identifies some other relevant notions of ‘directness’. For reference my ‘immediate perception’ is akin of his ‘perceptual directness’ and my ‘direct perception’ to his ‘epistemological directness’.

One might argue that if I have a perceptual experience as of a black cup of coffee before me, that experience is not sufficient to justify the belief that there is a black cup of coffee before me. This is because the experience could be veridical but could also be illusory or hallucinatory. If one accepts this line of thought, then it seems that perceptual experience alone is not sufficient to justify perceptual beliefs about ordinary physical objects. From here one might, for example, argue that the experience does justify the belief that there *seems* or *appears* to be a black cup of coffee before me. This latter belief would then be a direct (or basic) perceptual belief, and the former belief (that there is a black cup of coffee before me) would be an indirect (or non-basic) one. This general debate in perceptual epistemology is not my concern.

There is, however, a specific type of perceptual belief that is important to this discussion, namely *perceptual demonstrative beliefs*. These are beliefs involving a demonstrative that purport to refer to something that is (by hypothesis) currently perceived. To simplify, they are beliefs of the form ‘That is P’, where ‘that’ purportedly denotes some perceived object (or property or event or whatever) and ‘P’ is a property term of some sort. For example, the belief “That is blue” or “That is a laptop” which is held in response to a perception of something (and purports to be about that thing). When demonstratives are employed in this kind of way I will call them *perceptual demonstratives*. Perceptual demonstrative beliefs of this sort are particularly simple perceptual beliefs that rely quite heavily on what is perceived. They rely heavily on what is perceived because demonstratives are perhaps the most compelling example of a directly referential term, where the meaning or content of an employed demonstrative is directly and arguably wholly given by the referent.⁴ In this case the demonstrative is a perceptual demonstrative, and thus the intended referent is something purportedly perceived. The subject of the belief is in this way “directly given” by the perception. This makes simple perceptual demonstrative beliefs strong candidates for beliefs that can be directly and fully justified by perceptions – that is, they are candidates for one type of basic belief. While I am tempted by this view, for present purposes I will set aside debates about basic beliefs and focus on perceptual demonstratives and their associated beliefs as phenomena of interest in themselves. At minimum, interest in perceptual demonstratives stems from the fact that their successes seem to hinge precisely on the perceptions that they are parasitic on.

Most believe that we can demonstratively refer to ordinary physical objects like computers, cars, and trees. If a successfully employed perceptual demonstrative owes its success to the perception that it is parasitic on, then one might suppose that our account of perception, and in particular of immediate perception, should explain how we can perceptually demonstrate ordinary physical objects. Put another way, successful perceptual demonstratives of ordinary physical objects involve direct perception (in the epistemic sense), and this directness derives from the associated current perception. Perhaps we can and should explain this epistemic directness in terms of what one is immediately perceiving (in the perceptual sense). This is Bermúdez’s Reference Constraint:

Reference Constraint: If it is indeed the case that we make demonstrative reference to ordinary physical objects, then our account of the immediate object of perception must explain how this is possible. (Bermúdez, p. 365)

The Reference Constraint asserts that we should explain successful perceptual demonstratives in terms of immediate perception, setting the ground for explaining an important type of perceptual belief in terms of immediate perception. I will grant the Reference Constraint as a working assumption.⁵ The Reference Constraint prompts us to analyse immediate and mediate perception (the perceptual distinction). It also sets the stage for Bermúdez’s main argument. Let me turn to those now.

⁴ The character of the demonstrative, in Kaplan’s sense, is something like “the thing being demonstrated” or “the thing one is intending to demonstrate.”

⁵ Lyons (2023) provides an overview of perceptual epistemology that contains some options that may not require the Reference Constraint. I cannot delve into this terrain in the present work, though I will conclude that the Reference Constraint is important and worth preserving.

2. Bermúdez's main argument

Bermúdez considers two views of immediate perception: what we usually mean when we speak of sense-datum theory (SDT) and his proposed alternative, naturalized sense-datum theory (NSD). Stated most generally, the term 'sense-datum' means what is given to the senses. Studies of sense-data aim to give a theory of what it is that is given to the senses, of how what is given is related to the ordinary physical world and to ourselves, of how we should conceive of perception in this context, and of how all of this feeds into perceptual beliefs, knowledge and other aspects of epistemology. In this regard SDT and NSD are two theories about the nature of sense-data.⁶ Here are the core commitments of each.

According to NSD, "The immediate objects of visual perception, or what are often called sense-data, are parts of the facing surfaces of physical objects" (353). We nonetheless can and do perceive ordinary physical objects. However, according to NSD, we perceive ordinary physical objects by perceiving their front-facing surfaces. In this sense we mediate perceive those objects. The idea that we immediately perceive only the front-facing surfaces of objects is familiar from discussions of sense-data. I will unpack it below.

According to SDT, the immediate objects of perception are mind-dependent objects. Within SDT these mind-dependent objects are themselves typically called 'sense-data'. We thus have two uses of 'sense-data'. To help us keep track of this, I will use the term 'SD' for the mind-dependent objects that are central to SDT and continue to use 'sense-data' for the immediate objects of perception (whatever they may be). According to SDT, although the immediate objects of perception are these peculiar mind-dependent objects, namely SD, we can nonetheless perceive ordinary physical objects by immediately perceiving the mind-dependent ones. This is because SD *represent* ordinary physical objects, either intrinsically or in virtue of the way perceivers interpret them. In this way we mediate perceive ordinary physical objects. There are numerous arguments in favour of SDT, and at least as many criticisms. In my view the most powerful argument in favour of SDT stems from perceptual error, namely illusion and hallucination (see Section 6), and other important ones stem from causation and perception (see esp. Robinson, 1994), the scientific conception of the world in contrast to the phenomenal one (see esp. Jackson, 1977), and so on.

We thus see that, according to both NSD and SDT, we mediate perceive ordinary physical objects by immediately perceiving sense-data. However, what counts as sense-data (the immediate objects of perception) is strikingly different on the two accounts. According to NSD, sense-data are the front-facing surfaces of physical objects, and according to SDT sense-data are SD. We therefore mediate perceive physical objects for very different reasons on these views. This groundwork is sufficient to explain Bermúdez's overall argument, which I have distilled from Section III of his 2000 (pp. 365-372):

- (1) We do make demonstrative reference to ordinary physical objects.
- (2) The Reference Constraint is true.
- (3) There are only two viable theories of immediate perception: NSD and SDT.
- (4) SDT fails to adequately explain perceptual demonstratives about ordinary physical objects.
- (5) NSD successfully explains perceptual demonstratives about ordinary physical objects.
- (6) Therefore, we should endorse NSD (and reject SDT).

Premise (1) is put forward as a reasonable, intuitively justified hypothesis. One could object to it, including a defender of SDT. In this case the argument is stalled. However, I regard (1) as plausible and find the debate that ensues from as-

⁶ See Hatfield (2021) for an overview of sense-data. The history of the topic is complex, as is the relation between that history and the issues discussed in this work. For example, NSD bears an important relation to Moore's (1918-19) original version of sense-datum theory. What I will refer to as 'SDT' is the version that became more popular in subsequent decades, and is associated with Russell (1912), Jackson (1977), and Robinson (1994) among others. In the interests of space, I suppress these many important historical details.

suming its truth fascinating. Premise (2) is also a hypothesis, but it is justified by the kind of reasoning given in Section 1. Premise (3) is what Bermúdez takes to be the relevant views for this debate. I will remark on this in Section 8 but for now wish to note that even if one rejects the ‘only’ in premise (3), the argument could be modified so that (3) says that two of the important theories relevant to this debate are NSD and SDT, and the conclusion (6) asserts that NSD is preferable to SDT’. This would retain the validity of the overall argument, and (as I hope to show) retain the interest of the substance of Bermúdez’s insights. Premises (4) and (5) are clearly substantive claims in need of justification. Assuming they are justified, the argument is on solid footing.

At this point one might ask whether we need another argument against SDT, given how out of favour the view is in our current climate? I think this worry misses Bermúdez’s point and the interest of his paper. First, there is enduring interest in direct perception (epistemological notion) and perceptual demonstratives. In addition, the Reference Constraint is *prima facie* plausible and interesting. This prompts us to examine viable notions of immediate perception (and mediate perception) so that we can adequately understand how perceptual demonstratives can and by hypothesis do apply to ordinary physical objects. Second, the idea of trying to capture what discussions about sense-data were after (the immediate objects of perception, whatever they may be), put it to good use (via the Reference Constraint), and in the process show that SDT isn’t a good a way of doing this, is fascinating on its own. If Bermúdez’s argument works, then he effectively shows that SDT failed at its own game (giving a viable account of sense-data), but that we can nonetheless recover what the debate was after by embracing NSD. Third, the conceptions of immediate and mediate perception embodied in NSD and SDT are distinct, and stand on their own, independently of these theories. I will briefly remark on this in the conclusion, as I believe these disparate conceptions of immediate and mediate perception have fruitful, underexplored applications. Let me move on to justifying (5), and then critiquing (4).

3. *Why NSD?*

Bermúdez claims that (5) NSD successfully explains perceptual demonstratives about ordinary physical objects. I agree. Let me explain the idea. In NSD, we immediately perceive unoccluded front-facing surfaces of ordinary physical objects and mediate perceive the occluded parts of these objects and the objects as wholes. This is definitive of the view. To get a sense of the significance of the NSD, consider an everyday example and three ways of interpreting it.

Person Perception: When you look at a person who is facing you, there are parts of the person that are “presented” to you in a way that other parts are not. For example, their face and the front of their torso are oriented toward you and “in view” in way that their backsides, insides and body as a whole are not.

Consider three interpretations of this case.

One might propose an austere “parts only” theory of perception according to which all we really see are the unoccluded front-facing parts of physical objects. Judgements about whole objects and about occluded object parts are constructed in post-perceptual cognition, based on inferences from perceived front-facing parts of those objects. Thus, in Person Perception, you perceive only the front of the person’s face and torso. Insofar as you are “committed” to them having backsides and insides, and to them being a discrete whole physical object, such commitments arise only in post-perceptual cognitive judgement. This might be how to conceive of perceptual experience if something like Marr’s (1982) 2.5D sketch or Peacocke’s (1992) base level scenario content constituted perceptual experience.⁷

⁷ Neither Marr nor Peacocke restrict perceptual experience to the front facing parts of things. Instead, Marr posits the 2.5D sketch as present in an early stage of visual processing, and Peacocke posits scenario content as a key component of perceptual experience. Perceptual experience itself is for both theorists much richer than this.

Second, one might propose that we do experience whole objects, but that there is nothing to the idea that we “immediately perceive” the unoccluded front-facing parts or that there is any interesting difference between perceiving whole objects and perceiving their front-facing parts. This is an “objects only” view. Thus, in Person Perception, you perceive the person as a whole object and there is nothing distinctive about the difference between perceiving them as a whole object and perceiving their front-facing parts. On this view, we can safely set aside this whole discussion and talk simply about the objects that we perceive. This interpretation might be how to conceive of perceptual experience in contemporary naïve realism (e.g. Martin, 2017) or in some 4E theories of perception (e.g. Shapiro and Spaulding, 2024).

NSD proposes a different interpretation. According to NSD, and against the “objects only” view, there is an important sense in which we more immediately perceive the unoccluded front-facing parts of physical objects than their occluded parts, and than the objects as wholes. But NSD also asserts that we nonetheless generally perceive whole objects, in contrast to the “parts only” view. To hold these claims simultaneously, NSD proposes that we perceive all of these things but in different senses of ‘perceive’. Since our perceptions of the unoccluded front-facing parts of objects seem more immediate (or direct or present or epistemically available in perception), and our perceptions of whole objects and of their occluded parts occurs *by virtue of* our perceptions of the unoccluded front-facing parts of objects, the NSD theorist proposes that we mediate perceive whole physical objects and their occluded parts by immediately perceiving their unoccluded front-facing parts. Thus, in Person Perception, you immediately perceive the front of the person’s face and torso and mediate perceive them as whole objects and mediate perceive their occluded parts.

I take the NSD interpretation to be quite intuitive in part because it charts a well-motivated middle course between the more extreme “parts only” and “object only” views. Here are two further arguments for the approach. As Bermúdez argues (pp. 363–4), to perceive something we must discriminate it from its surrounds (see also e.g. Dretske, 1979). We generally visually discriminate objects by means of the front-facing parts that are open to view. Front-facing parts are thus serving a privileged perceptual role that is captured in the NSD conception of immediate perception and absent from the “objects only” view. Second, we can draw support for NSD from the phenomenon of amodal completion (see e.g. Briscoe, 2011). In psychology it is typical to distinguish between the front-facing parts of perceived objects and the occluded parts, and posit a special mechanism —amodal completion— to explain how we manage to perceive whole, complete objects despite in some sense only “immediately” perceiving their unoccluded front-facing parts. One important part of research in this area concerns illusions that involve amodal completion. Such illusions provide strong evidence for this kind of “filling in” being distinctive (against the “objects only” view), and being perceptual and not merely something arising in post-perceptual cognition (against the “parts only” view). Therefore, the NSD idea that we immediately perceive front-facing surfaces and mediate perceive ordinary physical objects is credible and a worthy foundation for a perceptual theory.⁸ In a broader discussion, I would delve deeper into this debate. However, for present purposes it suffices to situate NSD against some of its rivals, and explain why NSD advocates take the view to be compelling. Let me move to the substance of (5).

Premise (5) asserts that NSD can explain how we demonstrate ordinary physical objects (on the assumption that we do). The argument for (5) is as follows. According to NSD, we immediately perceive unoccluded front-facing parts of ordinary physical objects. It seems reasonable to assert that we can demonstrate what we immediately perceive: if there is anything that we can demonstrate, we can demonstrate what we can immediately perceive (in whatever sense of ‘immediately perceive’ one finds acceptable). One may deny this claim, for example by rejecting the very idea of immediate perception. However, I take it this wouldn’t be charitable to Bermúdez, and I’ve given adequate motivation for the existence of immediate perception in any case. One may also deny this by offering some theory of demonstratives that doesn’t rely heavily on what is perceived, but I will assume that perceptual demonstratives not only exist but are a foundational kind of demonstrative.⁹ Given the important link between demonstratives and perception, the claim that we

⁸ Note that I haven’t gone through Bermúdez’s potent analysis of some critiques of the NSD idea (see esp. pp. 359–364).

⁹ For comparison, Brovold and Grush argue that “the fashionable philosophical notion of the ‘perceptual demonstrative’ is an aberration” (2012, 39). There is a fascinating debate to be had here, but it falls outside my current purview.

can demonstrate what we immediately perceive is an attempt to carve out a basic and safe subset of perceptions and say that surely demonstratives, if they are applicable at all, must be applicable to this subset. I agree. It follows that we can demonstrate unoccluded front-facing surfaces of ordinary physical objects.

Bermúdez's key move at this point is to appeal to *deferred demonstration*. This is where a demonstration carries beyond what is before one in the following sense. For example, if one says "That man is the greatest!" while pointing to a picture of Messi, one is referring to Messi the person with 'that man' and not the picture. 'That' can thus go beyond a picture being perceived and refer to what the picture represents.¹⁰ To be sure, this doesn't prevent us from demonstrating the picture, for example by saying "That is a picture of the greatest footballer." Such a demonstration is *non-deferred* relative to the former demonstration.¹¹ Similarly, you might point to an oversized jacket on a chair and say "That's a large person" (deferred demonstration) or "That's a large jacket" (non-deferred demonstration). You may also hear a soundbite from an advert and exclaim "This is a catchy tune" (non-deferred demonstration) or "This is my favourite restaurant!" (deferred demonstration). I think that a full, adequate theory of deferred and non-deferred demonstration is difficult to develop, but the existence of the phenomenon is straightforward, and is appreciable independently of the issues we have discussed to this point.

Deferred demonstration is relevant in the following way. I have argued that we can demonstrate unoccluded front-facing surfaces of ordinary physical objects. These surfaces are parts of those objects. It seems plausible that deferred demonstration can "carry over the part-whole relation" (Bermúdez, p. 370). This is arguably true in general and in any case is very compelling when the part at issue is being immediately perceived and the whole at issue is being mediately perceived in the NSD sense of these terms. NSD therefore provides a compelling explanation of how demonstrative reference to ordinary physical objects can occur. I think we should accept this argument and hence (5).

4. *Why reject SDT?*

Premise (4) asserts that SDT cannot do what I, following Bermúdez, just argued NSD can, namely explain demonstrative reference to ordinary physical objects. Recall that according to SDT, we immediately perceive the mind-dependent objects SD, and by virtue of these representing ordinary physical objects (either intrinsically or by virtue of how we interpret them¹²) we mediately perceive the latter. Bermúdez's argument for (4) is to pose a dilemma for SDT. There are two possible ways to demonstratively refer to ordinary physical objects: via deferred and non-deferred demonstration. Neither model, he argues, can be applied within the constraints of SDT. Thus, (4) is true —SDT cannot explain demonstrative reference to ordinary physical objects. Let me unpack this.

Let's consider deferred demonstration first. Within SDT, we immediately perceive SD. Bermúdez argues that deferred demonstration to ordinary physical objects cannot be appealed to by SDT advocates. To see why, suppose that a perceiver immediately perceives a given SD_m ¹³, and mentally demonstrates that SD_m while saying or thinking "That is a laptop" with the intention of demonstrating the laptop that by hypothesis is represented by that SD_m . If this demonstrative succeeds in denoting the laptop, this would involve deferred demonstration. The reason Bermúdez believes this isn't credible is because:

linguists and philosophers of language are agreed that deferred ostension [demonstration] can only work in a communicative context when speaker and hearer share knowledge of an individuating principle linking the physically present ostended [demonstrated] object with the intended object of reference. (p. 372)

¹⁰ Bermúdez (370) helpfully reminds us that this idea is found in Quine (1968, pp. 40-44).

¹¹ I will stick with the terminology 'deferred' and 'non-deferred' when discussing demonstratives so as not to get confused with the other terminology already in place.

¹² I suppress this qualification below.

¹³ I will use ' SD_m ' to refer to a singular sense-datum, and continue to use 'SD' to refer to the plural sense-data.

However, Bermúdez argues, within SDT this requires shared knowledge of the speaker's private SD and the individuating principle that links those SD and ordinary physical objects. "It seems obvious, however, that no such principle is implicated in everyday demonstrative reference to material objects" (p. 372). Thus, Bermúdez concludes, SDT cannot appeal to deferred demonstration to explain demonstrative reference to material objects.

In other words, deferred demonstration works with a picture of Messi because I and my interlocutors understand the link (or individuating principle) between the picture and Messi and can implicitly appeal to this to correctly interpret what I'm trying—and by hypothesis succeeding—in saying when I utter "That is a great football player." By contrast, when I mentally point to my SD_m and say "That is a laptop", my interlocutors cannot have analogous knowledge of the requisite individuating principle because my SD_m and my demonstration of it are private. Thus, deferred demonstration is unavailable within SDT to explain demonstrative reference to ordinary physical objects.

This leaves SDT to explain demonstrative reference to ordinary physical objects via non-deferred demonstration. For Bermúdez this yields the other horn of the dilemma, for here demonstrative reference must somehow bypass SD and latch onto the ordinary physical object. The problem, however, is that this violates "the Reference Constraint because their account of the immediate object of perception has no part to play in their explanation of how demonstrative reference to three-dimensional material objects is achieved. The sense datum is otiose in the explanation of direct perception" (p. 372).

Bermúdez concludes that, whether SDT utilizes deferred or non-deferred demonstration, it cannot adequately explain how we demonstratively refer to ordinary physical objects. (4) is thus true.

5. *In defense of SDT*

I reject both horns of the dilemma Bermúdez tries to impose on SDT. Before addressing these claims directly, let me remark on two relevant issues about SDT. First, there is a general worry about SDT that stems from its appeal to a private relationship between a subject and her SD. One might hold that these kinds of private relations are problematic, perhaps by appeal to Wittgenstein's private language argument. I cannot engage in a detailed discussion of this topic. However, I can point to an important work like Fodor (1975) which provides a detailed account of how a private language of thought can coherently exist and arguably does exist in humans. This is sufficient to counter any blanket rejections of my proposal by appeal to worries about private languages. I therefore assume that private languages are possible. Among other things, this means that the Reference Constraint can in principle be interpreted to apply to private languages.

Second, one might worry that SDT leads to skepticism about the external world. How could we come to know anything about a mind-independent, physical world if all we immediately perceive are mind-dependent SD? Russell (1912) addressed this worry by appeal to the various patterns and stabilities that one can observe between SD. For example, every evening the SD I associate with my bed and its many features are like the ones I experienced when I left for work in the morning. Russell argued that the best explanation of the observed patterns and stabilities of an individual's SD is that there is a distal physical world that has similar patterns and stabilities and is the cause of one's SD. In principle, this reasoning is analogous to the reasoning that justifies our beliefs about things that we cannot "immediately observe" in nature, including subatomic particles and black holes. This includes justified beliefs about causal relations between those unobservables, and causal relations between those unobservables and our experiences of data observed on (e.g.) computer monitors that has been gathered by measurement devices like particle accelerators and the James Webb telescope. The full rational story of how we acquire and justify beliefs in unobservables in nature is complex, as is the full rational story of how, given SDT, one could come to believe in a distal physical world as the cause of one's SD. Regardless, I submit that SDT can avoid collapsing into a kind of idealism by an inference to the best explanation of what is the likely cause of various patterns of one's SD. I thus assume that SDT is consist-

ent with knowledge of a mediately perceived physical world.¹⁴ With these two assumptions in hand, let me return to Bermúdez's claims.

5.1. *The First horn*

The first horn of Bermúdez's dilemma concerns using deferred demonstration, where the SD are demonstrated and ordinary physical objects are referred to via deferred demonstration. Bermúdez's objection rests on the hypothesis that deferred demonstration requires shared knowledge by speaker and hearer of an individuating principle linking the demonstrated object (e.g. the picture or the SD_m) to the intended referent (e.g. Messi or the laptop). He then claims that such shared knowledge is "obviously" not present in everyday demonstrative reference to physical objects. The first claim should be rejected and the second claim, while true, doesn't prevent the advocate of SDT from giving an account of how we might—in principle—use deferred demonstration to demonstrate ordinary physical objects.

Bermúdez claims that shared knowledge of a principle that links the demonstrated object to the intended object is required for deferred demonstration. It is true that deferred demonstrations expressed via public languages rest on shared knowledge of this sort. It is also true that our actual practice of deferred demonstrations developed via shared knowledge of this sort. The relevant question, however, is a principled one: if SDT were true, could someone demonstrate one of their own SD in an effort to generate a deferred demonstration of an ordinary physical object, and succeed in referring to that object? Let me explain why we should answer positively.

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, anymore than my efforts to refer to Messi via a picture of him would fail. It does not matter whether there is a shared or public principle that links the demonstrated object to the intended object.

The other claim in the first horn of Bermúdez's dilemma is that everyday public demonstrative reference to physical objects does not seem to involve shared knowledge of deferred reference through SD. This is doubtless true. Regardless of the truth of SDT, common sense does not operate as though anything like SDT true, and thus doesn't embody shared knowledge of the kind of deferred reference to the robin egg described above. The same is true of linguistics, and thus we wouldn't expect linguistic analyses of demonstrative reference to generally reflect anything like the kind of private demonstration I am appealing to. The relevant question, however, is not how best to interpret everyday practice. The relevant question is whether deferred reference to the robin egg can, in principle, be shared among a community of speakers. I see no reason why it can't. Here is a rough sketch of how this could come about.

Suppose a group of people come to believe that SDT is true and they are in fact correct in this belief. Given my initial assumptions, this means that they also believe that there is a surrounding physical world, and that these physical things are the typical causes of each person's private SD. At the beginning, they may speak in a cumbersome technical way about SD and the physical world. For example, suppose A says:

"The physical thing that is causing me to have this blue, oval SD_m ..."

¹⁴ Russell is well known for having (1927) argued that knowledge of the physical world is limited to knowledge of the world's "structural" features. This limitation of worldly knowledge is sufficient for my purposes.

Here the demonstrative ‘this’ purports to refer to A’s SD_m . A community member B can understand this, even if, for example, B doesn’t rule out the skeptical possibility that A’s SD are (e.g.) inverted with respect to B’s. Furthermore, A can say:

“That robin egg is causing me to experience this blue, oval SD_m .”

Here, as before, ‘this’ demonstrates A’s SD_m and ‘that’ demonstrates the hypothesized egg. B has no difficulty understanding A’s claim. This would, for example, involve deferred demonstration of the egg if A’s employment of ‘that’ involves A mentally pointing to her SD_m with the intention of referring to the egg that A’s blue oval SD_m represents. If needed, A could explain to B that this is what she is trying to do, and then do it. B would thus understand that deferred demonstration was used to refer to the hypothesized egg with ‘that’ when A utters “That egg is causing me to experience this blue, oval SD_m .”

One may doubt whether B could come to know which physical object A is referring to—that is, can B come to know that A is talking about the specific robin egg in question? Yes. To see why, 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. While this scenario is underspecified in many ways, it strikes me as entirely coherent. Thus, while a shared principle of individuation of the right sort is required for shared communication of demonstrated objects, it is not required for deferred demonstration itself.

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 to, with ‘that’, refer to the hypothesized egg. This would also involve deferred demonstration of the sort that Bermúdez rejects, but the shared knowledge of the individuating principle would be implicit in their communicative practice, much as it is implicit in our practice when I point to the picture of Messi and say “That is a great football player.”

To be sure, Bermúdez’s objection is that it “seems obvious, however, that no such principle is implicated in everyday demonstrative reference to material objects” (p. 372). As a point about how to understand our everyday linguistic practice, I agree. The way we tend to understand our everyday actions is not in accordance with the SDT community just described. But this isn’t an objection to the ability of SDT to explain demonstrative reference to ordinary physical objects. Further, the SDT advocate can argue that, in the end, their description is *in fact* an accurate description of our everyday practice, whether we realize it or not. Regardless, I conclude that SDT can explain demonstrative reference to ordinary physical objects via deferred demonstration.

5.2. *The Second Horn*

The second horn of the dilemma asserts that SDT cannot use non-deferred demonstration to explain demonstrative reference to ordinary physical objects. This is because to do this, the demonstration would have to bypass the SD, in which case the demonstration doesn’t appeal to the immediate objects of perception (i.e. the SD), violating the Reference Constraint. I disagree. Return to speaker A in a community that believes SDT and in which SDT is true. Suppose A says “That egg is my favourite” and with ‘that’ succeeds in bypassing her blue, oval SD_m and directly referring to the purported egg. It doesn’t follow that the explanation of how this works can avoid appeal to A’s SD. Assume A’s demonstration consists entirely of her publicly pointing with her finger to the cup, and thus avoids the kind of inner demonstration by A to her own SD discussed in 5.1. Now suppose that A’s act of demonstrating occurs in significant part because of the immediate objects of her perception: she is able to perceptually connect to the egg, form her thought about the egg, and move her arm so as to point at the egg *because* of the SD_m that she is experiencing and the fact that the SD_m

represents the egg. On this account, A's demonstration of the egg is non-deferred in the relevant sense, but its success depends critically on the immediate objects of her perception, namely her SD.¹⁵ The Reference Constraint is therefore preserved. This strikes me as viable within SDT.¹⁶ This practice is also much closer to our everyday practice.

In conclusion, premise (4) of Bermúdez's overall argument, that SDT fails to adequately explain perceptual demonstratives about ordinary physical objects, should be rejected. Since this is Bermúdez's primary motive for rejecting SDT in favour of NSD, SDT can withstand his assault. However, there is much more of interest to be said about these issues. At minimum, we should consider whether NSD can explain the phenomena that originally motivated SDT (section 6), and, if not, whether it might be supplemented to do so (section 7).

6. *The problem of perception & NSD*

Suppose I am correct that Bermúdez has not given us a good reason to reject SDT. As mentioned earlier, the opponent of SDT may not be stressed, believing that there are several other reasons to reject SDT. Fair enough, and this is not the place to itemize and attempt to reply to each of them. However, I do wish to highlight another part of this dialectic. The SDT advocate needn't strictly play defense. Instead, they can go on offense and argue that NSD cannot adequately explain what is often called 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.¹⁷ Let me explain how The Problem of Perception is particularly challenging for NSD.

Here is a simple characterization of the core notions that is sufficient for our purposes.¹⁸ A veridical perceptual experience is fully accurate of the physical world. For example, you have an experience of a blue cylindrical thing, and you are in fact looking at a blue cylindrical thing. An illusion occurs when you experience a physical object, but experience at least one of its features incorrectly. For example, you have an experience of a blue cylindrical thing, and you are in fact looking at a purple cylindrical thing. An hallucination occurs when your experience doesn't have a corresponding physical object at all. For example, you have an experience of a blue cylindrical thing and there is nothing corresponding to this in the physical environment.

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. This explains why experiences of these types can be subjectively similar and even indistinguishable. However, the difference between veridical, illusory and hallucinatory experiences does not collapse, since which type of experience one is having depends on how well one's current SD latches onto (i.e. represents and is caused by) the physical world.

¹⁵ Compare with the "epistemic" role played by what is immediately perceived in Brown (2009, pp. 381-386).

¹⁶ One alternative possibility is to suppose that A demonstrates the cup solely by appeal to unconscious perceptual states. The perceptual experience would thus overdetermine what is needed for reference and by hypothesis for perceptual demonstrative belief. While this is possible, I do not believe this is an accurate model for understanding our perceptual lives (though I recognize that some, like Jack Lyons (personal communication), believe this model is largely correct). I also suspect that, to the extent that unconscious perception undermines SDT it also undermines NSD. These details, however, would take us too far afield.

¹⁷ The reasoning that leads from here to SDT was first clearly articulated by Robinson (1994) and has since been codified in Crane and French (2021). See again Hatfield (2021) for an overview of various additional motives for SDT.

¹⁸ For a powerful new theory of illusion and hallucination see Macpherson and Batty (2016).

It is straightforward that, on its own, NSD doesn't solve The Problem of Perception. NSD is a theory according to which the immediate objects of perception are the front-facing surfaces of ordinary physical objects, and ordinary objects themselves are mediate objects of perceptions. As it stands, this framework provides no resources to explain the difference between veridical perception, illusion and hallucination, nor the respect in which experiences of each sort can be subjectively indistinguishable. Bermúdez concedes one part of this challenge when he notes that NSD is "is only an account of non-hallucinatory perception" (p. 354). He thus recognizes that NSD needs to be bolstered to explain hallucination. However, he fails to make a similar concession about illusion, and about the subjective similarities that occur across experiences of all three sorts. He also fails to explicitly appreciate how acute the problem is.

Consider, briefly, illusion. 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). How should NSD be applied here? In line with NSD, the front-facing surface of the bottle is present before you in a way in which the backsides and insides are not and in which the bottle as a whole is not. However, the front-facing surface is purple, and you aren't experiencing that surface as purple, you are experiencing it as blue. In what sense, then, are you "immediately perceiving" the front-facing purple surface? If anything, it seems like the front-facing surface, although physically available, is perceptually occluded by the experienced blue. So, what is immediately perceived? Perceptually, it seems like a blue cylindrical bottle is immediately perceived, but there is nothing in the NSD view that captures this. If the NSD theorist adds something new to explain the illusion, whatever is "seemingly blue" arguably has to (a) capture the sense in which one is experiencing blue, (b) do so in a way such that blue is occluding the purple of the object, while (c) preserving the core NSD thesis that the front-facing surface of the perceived purple object is immediately perceived.

This is a problem for NSD because NSD is built around the idea of immediate perception and in particular the idea that front-facing surfaces are what is immediately perceived. There is no easy way for the NSD theorist to address these worries.¹⁹ To see why, consider a possible reply.

7. *Nonconceptual content & NSD*

Bermúdez is a well-known defender of nonconceptual content (e.g. Bermúdez, 1995, 2007; for an overview see Bermúdez & Cahen, 2024). It would be natural for him to appeal to nonconceptual content to address The Problem of Perception.²⁰ To what extent can nonconceptual content solve The Problem of Perception? How can this purported solution be integrated with NSD? These are the questions that I address in this section. My conclusions are as follows. Nonconceptual content can adequately address the semantic part of The Problem of Perception. However, it can only positively address the phenomenological part by appeal to postulates that are no less extreme than those of SDT. Further, the resulting framework is either incoherent, or asserts that in veridical perceptual we are mediately aware of the physical world. The latter commits NSD to a perceptual veil that is similar to the one embraced by SDT, and which NSD is designed to avoid. In effect, via nonconceptual content NSD can solve the Problem of Perception, but only by embracing a veil of perception. Let me unpack this more slowly.

A full discussion of nonconceptual content falls outside the scope of this work. For our purposes, it is instructive to recall what nonconceptual content is and what kinds of reasons are offered for its existence. Contents are expressed by

¹⁹ Compare with Hatfield's (2021) worry that NSD may have to postulate "appearances" to accommodate perceptual constancy.

²⁰ Thank you to an anonymous referee for suggesting that I address this possibility.

representations.²¹ Nonconceptual content is a form of content that is in principle distinct from conceptual content, that is, distinct from contents found in thought. Bermúdez’s “Master Argument” for nonconceptual content in perception rests on two principles:

1. “[T]he content of perception is linked to capacities for perceptual discrimination”;
2. “[C]oncept possession can never wholly be explained in terms of capacities for perceptual discrimination.” (Bermúdez, 2007, 59)

From these principles it follows that there is perceptual content that is nonconceptual.

There are important case studies that help justify this conclusion. Many animals can perceptually discriminate things that their cognitive systems are not sufficiently complex to generate concepts for. This point carries over to humans as well. We can perceptually discriminate detailed, nuanced features like highly specific colours and shapes despite not having concepts for such specific properties. These cases suggest that perceptual discrimination is operating, in a fundamental way, independently of cognition, and thus that perceptual contents are nonconceptual.

Contents, by their nature, can have varying degrees of accuracy to their subject. Thus, Bermúdez identifies the possibility of misrepresentation as one of the core criteria of representations (1995, p. 351). Attributing nonconceptual content to perceptual experience thus permits, indeed guarantees, the possibility of perceptual error. This makes it poised to help address The Problem of Perception.²²

Recall that, as standardly conceived, in illusion there is successful object perception but erroneous property perception. For example, one experiences a purple bottle as blue. By contrast, successful object perception does not occur in hallucinations. For example, one has an experience as of purple bottle and there is nothing in the physical world that corresponds to this. Let me unpack what a theory of perceptual error must explain, focusing again on illusion. The key distinction for our discussion is between the *semantic* and *phenomenological* challenges posed by perceptual error.

One fundamental puzzle raised by illusion is a *semantic* one of how to model perceptual experience such that it can be said to involve the perception of an object but the misperception of one or more of its features. It is straightforward that adding nonconceptual content to NSD affords an account of the semantic challenge of illusion that is unavailable to NSD proper. In the bottle case, I am perceiving a bottle that is in fact purple but erroneously experiencing it as blue. By appeal to nonconceptual content, we can assert, for example, that my perceptual system is expressing something like the content <there is a blue bottle>²³. Thus, my perceptual system is correctly representing that there is a bottle but incorrectly representing it as blue. For our purposes, this is an adequate explanation of the semantic challenge of illusion. However, it is important to see that this account of illusion is incomplete, and incomplete in a manner that is at the heart of our discussion.

It is one thing to give a model of the semantics of illusion. It is quite another to give a model of the *phenomenology* of illusion, a model of what it is like for the perceiver to undergo an illusion. On its own, nonconceptual content yields no positive account of the phenomenology of illusion. One way to see this is to recognize that nonconceptual content is consistent with SDT. As I’ve emphasized throughout, according to SDT, SD represent the physical world, either intrinsically or via an interpretation of SD. Either way, this is consistent with SD expressing nonconceptual content. For example, a charitable interpretation of SDT holds that humans can perceive many specific colours and shapes in

²¹ Some of the terminology that follows I will employ without argument (e.g. the term ‘expressed’ in above sentence). Hopefully, context is sufficient to capture what is intended. In all cases I intend the terminology to be as neutral as possible.

²² What follows is a discussion of a contemporary form of representationalism about perceptual experience. As is well known, there are now numerous approaches to representationalism. I cannot survey them all, but what follows is representative of a very influential approach.

²³ There are of course various ways to unpack the semantics of perceptual contents. The existential format in the text is reminiscent of Tye (1995, 2002). These details don’t substantively impact the argument that follows.

the physical world that they don't have concepts for *because* SD can represent specific colours and shapes that humans don't have concepts for. This permits humans to mediately perceive those specific physical features. The same holds for animals who can discriminate things they don't possess concepts for.

Regarding illusion, SDT already embodies the semantic solution offered by nonconceptual content: when one misperceives the purple bottle as blue, one's blue, bottle-shaped SD correctly represents that there is a bottle in the physical world but incorrectly represents it as blue. What SDT adds to the semantic account of illusion afforded by nonconceptual content is an account of the perceptual phenomenology: the bottle perceptually seems blue (in the phenomenal sense of 'seems') because one is *immediately perceiving* a blue bottle-shaped SD. This is a contentious commitment of SDT. It is also a positive account of the phenomenal challenge posed by illusion and, by extension, by hallucination. What can the NSD advocate say?

On its own, appeal to nonconceptual content does not provide a positive account of the phenomenology of perceptual error. I, for one, cannot find such a positive account in Bermúdez's work. Thus, to provide a concrete analysis I will make two simplifying, charitable assumptions. First, other well-known defenders of nonconceptual content have offered substantive and similar proposals. Here are two examples:

Along with (most) other representationalists, I am happy to say that, in the hallucinatory case, the perceiver is conscious of an un-instantiated property. (Tye, 2014, p. 304, fn. 20)

In hallucinating...We are aware of pure universals, uninstantiated properties (Dretske, 2003, p. 73)

Dretske and Tye postulate uninstantiated properties (e.g. Platonic universals²⁴) and a capacity to be perceptually conscious or aware of them. These tools are then used to explain the phenomenology of hallucination. For ease of reference call this first assumption *the Dretske-Tye account of hallucination*. This proposal has the advantage of providing a positive account of the phenomenological puzzle. However, it raises two challenges. It relies on metaphysical entities that are not straightforwardly friendly to naturalism (e.g. universals are generally regarded as being outside space and time) and it explains perceptual phenomenology via a novel perceptual capacity to become perceptually aware of the proposed metaphysical entities. More concretely, it explains perceptual phenomenology via awareness of entities outside space and time.²⁵ One reason this explanation is questionable is because there is a decided "here and now" aspect to perceptual experience that seems at odds with universals. One would not blame the SDT advocate for arguing that explaining the phenomenology of hallucinations via a capacity to immediately perceive SD is no more troubling than explaining it via awareness of universals outside space and time.²⁶

Let me first conclude that (a) there are influential defenders of nonconceptual content who recognize a need to provide not only a semantic account of perceptual error but also a phenomenological one, and (b) their positive phenomenological account appeals to postulates that are, on their surface, no less worrying than the postulates of SDT.

To extend this account to veridical perception and illusion we need to make another assumption. For ease of discussion, I will confine myself to *common-factor views*, which are arguably endorsed by Bermúdez, Dretske, and Tye. According to common-factor views, the nature of perceptual experience is fundamentally the same across hallucination, illusion and veridical perception —what varies is how those experiences connect to (i.e. causally and semantically) the physical world.²⁷ As mentioned, SDT is a common-factor view. Applying the common-factor view to the Dretske-Tye

²⁴ While Dretske appeals to universals to understand uninstantiated properties, Tye (2014) appeals to sets of possible worlds. For simplicity, I will stick to universals in the text. I don't believe this affects the substance of my argument.

²⁵ In the case of Tye's preferred metaphysical postulated, sets of possible worlds, the result is that perceptual phenomenology is explained via awareness of sets possible worlds, i.e. awareness of a collection of entities that constitute different spacetimes.

²⁶ Compare with chapter 5 of Schellenberg (2018).

²⁷ By contrast, disjunctivist views hold that the nature of perceptual experience is fundamentally different, depending on whether the experience is veridical or erroneous. Disjunctivism is standardly aligned with nonrepresentational views of perception, though there are excep-

account of hallucination arguably yields the conclusion that one is aware of universals during hallucinations, illusions and veridical perceptions. What varies is how those experiences causally and semantically connect to the physical world.

One odd consequence of this view is that it isn't explicit if or when we are aware of the physical world. Instead, to this point, awareness is a relation we bear to universals, and the physical world is only entering the theory at the semantic and causal levels (i.e. where one explains how well the physical world satisfies a given experience/ content). Yet, surely, we are aware of the physical world in some sense. One natural suggestion is that our awareness of the physical world, in some sense, is mediated by our awareness of universals. That is, if experience has the same nature across hallucinations, illusion and veridical perceptions, and involves awareness of universals in all cases, then, when one is aware of the physical world in perceptual experience, such awareness seems to be mediated by one's awareness of universals. One's awareness of the physical world arises through that world satisfying the structure of universals that explains the phenomenology of one's experience in a given case. For example, if one has an experience as of a blue bottle, then one is aware of the structured universal <there is a blue, bottle-shaped object>. In the veridical case, to the extent that one is aware of the relevant blue bottle, one is aware of it by virtue of that bottle satisfying this structure of universals.

The problem is that this is starting to sound suspiciously like the framework espoused by SDT. While this doesn't commit to the *immediate perception* of universals and the *mediate perception* of the physical world, it is committed to the *immediate awareness* of universals and the *mediate awareness* of the physical world, in a sense of these terms that looks very friendly to the SDT framework. Suppose, for example, that SDT were altered so that, instead of perceivers immediately perceiving SD, perceivers are immediately *aware* of SD, and instead of perceivers mediate perceiving the physical world, perceivers are mediate *aware* of the physical world. I am not sure what of substance this would change in SDT. In this case, both the revised SDT and the generalized Dretske-Tye views would then be committed to the thesis that perceivers are immediately *aware* of some entity type X and mediate *aware* of the physical world (where, for all x, x ≠ the physical world). Put another way, to explain The Problem of Perception the union of NSD and nonconceptual content is committed to a distinction between immediate and mediate perceptual consciousness of a sort that NSD was designed to avoid, i.e. one that is akin to the sort utilized in SDT. That is my second claim.²⁸

My third claim, which is a corollary of the second, is that the case of illusion brings out this problem in a particularly forceful way. I sketched this idea in the last section, but can state the problem with more precision given the resources that are now available. Illusions are challenging because they are semantically-hybrid states involving successful object perception and at least one unsuccessful property perception. When I experience the purple bottle as a blue bottle, I am successfully perceiving the bottle and the bottle shape but not successfully perceiving the bottle colour. Pretheoretically, the purple of the bottle is occluded by the experience of blue (however one wants to unpack the latter). According to NSD, during perceptual success I immediately perceive the front-facing surfaces of objects, and mediate perceive the objects themselves. Given that I am successfully perceiving the bottle, its shape and its location, it seems to follow (within NSD) that I am immediately perceiving the front-facing surface of the bottle. However, given that the purple of the bottle is perceptually occluded, and that the bottle's front facing surface is imbued with that purple, it seems hard to assert that I am immediately perceiving the front-facing surface of the bottle. This is a problem for NSD, even when supplemented with nonconceptual content.

I see no compelling way to resolve this tension without collapsing NSD into an SDT-friendly view. According to the solution outlined above, although I am immediately perceiving the front-facing surface of the bottle in the NSD sense

tions (e.g. Johnston, 2004; Schellenberg, 2018). Crane and French (2021) provide an overview. For the interested reader, Tye (2014) critiques so-called "gappy content" views (which e.g. Schellenberg defends) and defends two niche common-factor content views, surveying numerous alternatives in his discussion.

²⁸ It is instructive to compare the above argument with that found in Kriegel (2011). Like me, Kriegel argues that views of the sort just discussed are committed to a veil of abstracta. However, he then argues that this must be avoided, and can be avoided by adverbialism, concluding that we should adopt adverbialism. I, unsurprisingly, reject the second stage of his reasoning. In the interest of space and focus, I cannot work through the details here.

of the term, I am nonetheless mediately aware of that surface in the sense that my awareness of the surface is mediated by my awareness of universals. That is, NSD seems to collapse into a view akin to SDT in the sense that NSD is committed to the claim that we are not immediately aware of the physical world, even during veridical perception.

In conclusion, this argument rests on two assumptions, the Dretske-Tye account of the phenomenology of hallucination, and a common-factor account of nonconceptual content. These assumptions are justified because they are defended by leading advocates of nonconceptual content. Nonetheless, the NSD advocate is welcome to suggest alternative assumptions.²⁹ For our purposes, I submit that I've adequately justified my conclusion: when extended to the Problem of Perception via an appeal to nonconceptual content, NSD is either rendered incoherent or collapses into a view that is akin to SDT in that both views deny that we are immediately aware of the physical world in perceptual experience.

8. *Is it SDT, NSD or bust?*

NSD and SDT are viable views, despite facing some challenges. As I hope to have shown, examining how they conceive of immediate and mediate perception, and how they explain demonstrative reference to ordinary physical objects, is incredibly instructive. Nonetheless, one might wonder whether we should embrace premise (3), that NSD and SDT are the only two viable theories of immediate perception. It is thus worth briefly considering an alternative that has received recent attention: naïve realism.

According to this view, we immediately perceive not the front-facing surfaces of physical objects but physical objects themselves. That is, naïve realism is built around the idea that we immediately perceive ordinary objects like trees and cars. In this regard the view differs from NSD. However, Bermúdez dismisses naïve realism because it is “no longer in play” (p. 368). While that was true in 2000, when Bermúdez's article was published, in our current climate naïve realism is very much in play.³⁰ Naïve realism satisfies the Reference Constraint, as the view can explain perceptual demonstratives by the immediate perception of ordinary objects. In this regard, were one to defend NSD in today's climate, one would arguably be obliged to provide reason to prefer NSD to naïve realism.

A defense of NSD over naïve realism can begin with the arguments NSD advocates offer in favour of treating the front-facing surfaces of physical objects, and not the whole objects themselves, as the immediate objects of perception (Section 3). Recall, one argument for NSD asserts that we visually discriminate objects by their front-facing surfaces and thus perceive front-facing surface more immediately than we perceive objects themselves (Bermúdez, 2000, pp. 363–4). The other argument appeals to the psychological capacity for amodal completion (e.g. Briscoe 2011), which arguably entails that there is a primacy to the perception of the front-facing surfaces of objects over the perception of objects themselves.

There is much interesting work to do here, but this work sadly falls outside my present purview. I will emphasize that, however the debate between NSD and naïve realism is resolved, the challenge from SDT remains. SDT theorists demand an explanation of The Problem of Perception. Providing such an explanation is not only a difficulty for NSD, but also a well-known difficulty for naïve realism.³¹ In this regard SDT again sits in a comfortable position.

²⁹ For example, Pautz (2021) defends a quietist account of perceptual phenomenology. According to his view, in erroneous perceptual experience it indeed seems like one is aware of uninstantiated properties, but no positive account of this is needed. Why? Roughly, because a quietist view of this sort is superior to all non-quietist alternatives. To me, as a solution this has all the advantages of theft over honest toil.

³⁰ In fact, the recent interest in naïve realism arguably began very soon after Bermúdez's (2000). For example, two influential works came out in 2002: Campbell (2002) and Martin (2002). See Crane and French (2021) for an overview.

³¹ Naïve realists have offered numerous proposals for how they might account for illusions. See e.g. Brewer (2008), Genone (2014) and French & Phillips (2020). See Brown & Macpherson (2026) for a critical discussion.

There are, of course, other options that should be considered in a wider discussion. For example, a distinctive, recent account is Hill's (2022). On Hill's view, we do not directly experience objective properties of physical objects (like their colour and size). Instead, a special class of abstract properties —Thouless properties— are the “most immediate objects” of perceptual experience (p. 50).³² Thus, Hill argues that “perceptual experience presents us with Thouless properties rather than their objective counterparts. This undercuts naïve or direct realism” (p. 57). This arguably also undercuts NSD and SDT.

There is thus much more to be said about premise (3). Nonetheless, even restricting our discussion to NSD and SDT (as premise (3) does) yields enormous fruit that can be applied to a broader suite of options in future work.

9. *The underlying framework, without SDT and NSD*

SDT and NSD appeal to different conceptions of the distinction between immediate and mediate perception. What has not been made explicit is that these conceptions do not depend on either theory. Here is each conception on its own:

- **Front-Facing Surface Conception:** the immediate₁ object of perception is a (front-facing) part of the mediate₁ object of perception. For example, you mediate₁ perceive the car by immediately₁ perceiving the front side of it.
- **Representation Conception:** the immediate₂ object of perception represents (and is distinct from) the mediate₂ object of perception. For example, you mediate₂ perceive a football match by immediately₂ perceiving a television.

Although NSD is built around the first conception and SDT around the second, each conception can stand on its own. Further, each can and should play an important explanatory role in perceptual theory. Let me briefly explain why.

Regarding the Front-Facing Surface Conception, it is hard to deny that we have more immediate perceptual access to the front-facing parts of things than we do to their backsides, insides or to the things as wholes. The difference can, for example, be sidelined in psychophysical studies of two-dimensional figures on computer monitors. But this restriction is obviously an artefact of experimental design, and ultimately three-dimensional physical objects are a, if not the, fundamental topic of perceptual psychology and of everyday perception. With three-dimensional object perception comes the Front-Facing Surface Conception of immediate and mediate perception. We thus can and should debate about how to understand the Front-Facing Surface Conception and its significance for issues like object perception and perceptual demonstratives. However, at the end of this debate some form of this conception of immediate and mediate perception will remain intact.

Assuming SDT is false, the Representation Conception arguably played a small role in our evolutionary history. Independently of this, it is playing a decidedly outsized role in our current lives, with the proliferation of televisions, smartphones, computers, and the like. It is straightforward to me that my son is not merely growing up with a blurred understanding of the reality of his immediate physical environment versus the reality represented by various screens to which he has access. He is arguably growing up without a sharp division between these realities. I won't pause to try to assess the value of this. My point is that what is mediate₂ perceived is arguably as real to him as what is immediately₂ perceived. Our theories of perception should pay much closer attention to the Representation Conception than they have.

³² Thouless properties are a kind of appearance property that is postulated to explain deviations from perfect perceptual constancy found in vision. More precisely, regarding size, “The Thouless size of an object x with respect to viewpoint y is $F(v, d)$, where v is the visual angle x subtends with respect to y and d is information pertinent to the distance from x to y . The function F is a “partial-constancy” function, in the sense that its values are more stable than ever-fluctuating retinal images but are nonetheless always characterized by under-constancy” (2022, p. 45).

This is additionally true given the rapid development of virtual and augmented reality technology. The kinds of experiences found in these settings test the limits of our understanding of what is “immediately” versus “mediately” perceived, and of what we are referring to when we think and talk about what is experienced. The solution isn’t to abandon the topic of immediate and mediate perception, it is to develop it in a way that permits it to illuminate our burgeoning perceptual reality.

It is also worth making explicit that both the Representation Conception and the Front-Facing Surface Conception can be applied to a single perceptual experience. For example, one can immediately₁ perceive a front-facing surface of an object, and mediate₁ perceive the object itself, and, simultaneously, if that object represents some further object, then one can mediate₂ perceive the further object by immediately₂ perceiving the former. In my view this is precisely what happens when one is watching a live football match on a smartphone.

Above I argued that both deferred and non-deferred demonstration can be applied, in various ways, to cases to which the Front-Facing Surface Conception can be applied, and to cases to which the Representation Conception can be applied. A full account of these matters should appeal to all of these resources, regardless of whether one adheres to SDT, NSD or some other view. We need both conceptions of immediate and mediate perception, we need both deferred and non-deferred perceptual demonstratives, and, although the details are complex, many instances of the latter should be explained via combinations of the former. In this regard the Reference Constraint, or something in the vicinity, remains in place. I therefore see the underlying conceptual framework found in Bermúdez’s ‘Naturalized Sense Data’, when properly understood, to hold enduring value.

10. Conclusion

Throughout the twentieth century, philosophers of perception often discussed how immediately we perceive ordinary physical objects, and, if not immediately, then how what is immediately perceived feeds into and justifies basic perceptual beliefs about ordinary physical objects. In recent years, philosophy of perception has moved away from this suite of issues. In this regard, Bermúdez’s ‘Naturalized Sense Data’ may seem old fashioned. To my mind, viewing it as such is a mistake. There are different and legitimate conceptions of the distinction between mediate and immediate perception, and it is instructive to examine how each feeds into perceptual beliefs, such as those involving perceptual demonstratives about ordinary physical objects, regardless of whether one views these as basic beliefs. I argued that Bermúdez was wrong to contend that SDT could not adequately explain such demonstratives. Instead, I argued, there are multiple theories that can explain perceptual demonstratives. His preferred view, NSD, would also benefit from an account of The Problem of Perception: perceptual errors and their relation to veridical perceptions have always been important to the distinction between mediate and immediate perception. NSD is right to apply the distinction to veridical perceptions, but wrong to not also address perceptual error, and in particular illusion, given the forceful challenge illusion poses to the view. We can supplement NSD with nonconceptual content. However, in so doing we must be careful to not merely address the semantic component of The Problem of Perception. Addressing the phenomenological component requires an explicit account of perceptual phenomenology. I considered a proposal of Dretske and Tye, influential defenders of nonconceptual content. I argued that, in the end, the union of NSD with nonconceptual content yields either an incoherent view, or a view according to which we are not immediately aware of the physical world during veridical perceptions, but instead mediate₁ aware of it. This introduces a kind of perceptual veil that is akin to that embraced by SDT—an outcome Bermúdez sought to avoid.

In conclusion, I suggest that the underlying framework Bermúdez articulates to guide his discussion is essential to perceptual theory, regardless of which specific theory you endorse. His distinction between immediate/mediate perception and direct/indirect perceptual beliefs is critical (see also Jackson, 1977; Snowdon, 1992). Further, his distinction between two types of immediate-mediate perception—one embodied in what I’ve called SDT (reminiscent of Russell,

1912) and one in his NSD (reminiscent of Moore, 1918-1919)— is under-appreciated and, to my mind, foundational to perceptual theory. These distinctions can and should be abstracted from SDT and NSD and applied to numerous case studies and other perceptual theories. That landscape is at present largely unexplored. I hope it won't be for long.³³

Acknowledgements

This paper was written for the 30th SIUCC conference in honour of the work of José Luis Bermúdez, held in Valencia from 26-28 April 2023. I wish to thank the audience for helpful comments, and the conference organizers Victor Verdejo and Matheus Valente. Thanks are also owed to my colleagues David Bain, Robert Cowan, Jack Lyons and Fiona Macpherson, all of whom facilitated a discussion on this material that resulted in a dramatic improvement in my thinking. Thanks are also due to Robert Briscoe. Finally, thanks to the audience at the University of St. Andrews Philosophical Society. This work was supported by the Arts and Humanities Research Council AH/Y007638/1.

REFERENCES

- Bermúdez, J.L. (1995). Nonconceptual content: from perceptual experience to subpersonal computational states. *Mind and Language* 10 (4), 333-369.
- Bermúdez, J.L. (2000). Naturalized sense data. *Philosophy and Phenomenological Research* 61(2), 353-374.
- Bermúdez, J.L. (2007). What is at stake in the debate on nonconceptual content? *Philosophical Perspectives* 21 (1), 55-72.
- Bermúdez, J. & Cahen, A. (2024). Nonconceptual mental content. In Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition). URL = <<https://plato.stanford.edu/archives/fall2024/entries/content-nonconceptual/>>.
- Brewer, B. (2008). How to account for illusion. In A. Haddock and F. Macpherson (eds.) *Disjunctivism: Perception, action, knowledge* (pp. 168-80). Oxford University Press.
- Briscoe, R. E. (2011). Mental imagery and the varieties of amodal perception. *Pacific Philosophical Quarterly* 92 (2), 153-173.
- Brovold A. & R. Grush. (2012). Towards an (improved) interdisciplinary investigation of demonstrative reference. In A. Raftopoulos & P. Machamer (eds.), *Perception, realism, and the problem of reference* (pp. 11-42). Cambridge University Press.
- Brown, D. H. (2009). Indirect perceptual realism and demonstratives. *Philosophical Studies*, 145, 377-94.
- Brown, D. H. & Macpherson, F. (2026). *Colour illusion*. Routledge.
- Campbell, J. (2002). *Reference and consciousness*. Oxford University Press.
- Crane, T. & French, C. (2021). The problem of perception. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/fall2021/entries/perception-problem/>>.

³³ Project SENSOR (2024-2027), an AHRC-DFG funded project on “sensory engineering” has this as one of its main topics. Sensory engineering refers to ways perceptual experiences can be intentionally manipulated —engineered— via technologies of different sorts. See <https://www.gla.ac.uk/research/az/cspe/projects/sensorsensoryengineering/>.

- Dretske, F. (1979). Simple seeing. In Gustafson, D.F., Tapscott, B.L. (eds) *Body, mind, and method. Synthese Library*, 138 (pp. 1-15). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-9479-9_1
- Dretske, F. (2003). Experience as Representation. *Philosophical Issues* 13, 67-82.
- French, C. & Phillips, I. (2020). Austerity and illusion. *Philosophers' Imprint*, 20(15), 1-19.
- Fodor, J. (1975). *The language of thought*. Harvard University Press.
- Genone, J. (2014). Appearance and illusion. *Mind* 123(490), 339–376.
- Hatfield, G. (2021). Sense data. In Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). URL = <<https://plato.stanford.edu/archives/fall2021/entries/sense-data/>>.
- Hill, C. (2022). *Perceptual experience*. Oxford University Press.
- Jackson, F. (1977). *Perception*. Cambridge University Press.
- Johnston, J. (2004). The obscure object of hallucination. *Philosophical Studies*, 120(1), 113-183.
- Kriegel, U. (2011). The veil of abstracta. *Philosophical Issues* 21(1), 245-267.
- Lyons, J. (2023). Epistemological Problems of Perception. In Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). URL = <<https://plato.stanford.edu/archives/sum2023/entries/perception-episprob/>>.
- Macpherson, F & Batty, C. (2016). Redefining illusion and hallucination in light of new cases. *Philosophical Issues* 26(1), 263-296.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Martin, M.G.F. (2002). The transparency of experience. *Mind and Language*, 17, 376-425.
- Martin, M.G.F. (2017). Elusive objects. *Topoi* 36(2), 247-271.
- Moore, G. E. (1918-1919). Some judgements of perception. *Proceedings of the Aristotelian Society*, 19, 1-29.
- Pautz, A. (2021). *Perception*. Routledge.
- Peacocke, C. (1992). *A study of concepts*. MIT Press.
- Quine, W.V.O. (1968). Ontological relativity. In W.V.O. Quine (ed.) *Ontological relativity and other essays* (pp. 26-68). Columbia University Press, 1969.
- Robinson, H. (1994). *Perception*. Routledge.
- Russell, B. (1912). *The problems of philosophy*. Williams and Norgate.
- Russell, B. (1927). *Analysis of matter*. Kegan Paul.
- Schellenberg, S. (2018). *The unity of perception: content, consciousness, evidence*. Oxford University Press.
- Shapiro, L. & Spaulding, S. (2024). Embodied cognition. In Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition). URL = <<https://plato.stanford.edu/archives/fall2024/entries/embodied-cognition/>>.
- Snowdon, P. (1992). How to interpret 'direct perception'. In T. Crane (ed.), *The contents of experience* (pp. 48-78). Oxford University Press.

Tye, M. (1995). *Ten problems of consciousness*. MIT Press.

Tye, M. (2002). *Consciousness, color, and content*. MIT Press.

Tye, M. (2014). What is the content of a hallucinatory experience? In Berit Brogaard (ed.), *Does perception have content?* (pp. 291-330). Oxford University Press.

PROFESSOR DEREK H. BROWN worked at Brandon University (Canada) prior to joining the Centre for the Study of Perceptual Experience and Philosophy at University of Glasgow in 2017. He works on philosophy of perception, with particular interest in philosophy of colour and mediated perception. He is currently the UK-PI of the AHRC-DFG funded project *Sensory Engineering* (2024-2027), which is examining ways that sensory experiences can be intentionally manipulated via technologies.

ADDRESS: School of Humanities, University of Glasgow, G12 8QQ, UK. E-mail: derek.brown.2@glasgow.ac.uk – ORCID: <https://orcid.org/0000-0002-5765-2443>



WHY A PERCEPTUAL ZERO-POINT AMOUNTS TO NOTHING

(Por qué un punto cero perceptivo no significa nada)

Shaun Gallagher*

University of Memphis

University of Wollongong

<https://orcid.org/0000-0002-3147-9929>

Keywords

Zero-point

Perception

Action

Egocentric spatial frame of reference

ABSTRACT: The concept of a “zero-point,” the bodily origin of the first-person perspective or the egocentric spatial frame of reference through which we encounter the world, is pervasive in phenomenological accounts of perception. I will argue that the zero-point is a complete abstraction that misleads the analysis of perception and action. In contrast, body-schematic processes involved in perception and action are multi-dimensional, complex and complicated by temporal and intermodal factors. I consider several different characterizations of the zero-point: as an objective physiological point, as a geometrical point, or, phenomenologically, as part of one’s experience. I examine evidence from vision studies, and more generally from studies of sensory-motor processes, action, and intersubjective interaction. I reference the work of José Bermúdez in this area and briefly outline an alternative notion that better captures what the zero-point was meant to capture. Bermúdez’s analysis helps us to see that there is no point to the concept of zero-point.

Palabras clave

Punto cero

Percepción

Acción

Marco de referencia espacial egocéntrico

RESUMEN: El concepto de un «punto cero», el origen corporal de la perspectiva de primera persona, o del marco de referencia espacial egocéntrico desde el que nos relacionamos con el mundo, es recurrente en los enfoques fenomenológicos de la percepción. En este trabajo, argumentaré que el punto cero es una completa abstracción que distorsiona el análisis de la percepción y la acción. Por el contrario, los procesos esquemáticos corporales involucrados en la percepción y la acción son multidimensionales, complejos y están influidos por factores temporales e intermodales. Consideraré diferentes caracterizaciones del punto cero: como un punto fisiológico objetivo, como un punto geométrico, o, fenomenológicamente, como parte de la experiencia subjetiva. Examinaré evidencia proveniente de los estudios de la visión y, de forma más general, de estudio de los procesos sensoriomotores, la acción y las interacciones intersubjetivas. Haré referencia al trabajo de José Bermúdez en este campo y esbozaré brevemente una noción alternativa que capta de mejor manera lo que el concepto de punto cero pretendía capturar. El análisis de Bermúdez nos ayuda a ver que no hay razón para mantener el concepto de punto cero.

Gako-hitzak

Zero puntua

Pertzepzioa

Ekintza

Erreferentzia-marko espazial egozentrikoa

LABURPENA: «Zero puntu»aren kontzeptua —lehen pertsonako perspektibaren gorputz-jatorria edo munduari aurre egiteko baliatzen dugun erreferentzia-marko espazial egozentrikoa— oso presente dago pertzepzioaren deskribapen fenomenologikoetan. Ikusiko dugunez, zero puntua erabateko abstrakzioa da, eta pertzepzioaren eta ekintza-aren analisia desitxuratzen du. Aldiz, pertzepzioaren eta ekintza-aren eragina duten gorputz-prozesu eskematikoak multidimentsionalak eta konplexuak dira, eta faktore intermodal eta denborazkoek eragiten diete. Zero puntuaren hainbat karakterizazio aztertuko ditut: puntu fisiologiko objektibo gisa, puntu geometriko gisa edo, ikuspegi fenomenologikotik, norberaren esperientziaren parte gisa. Ikusmenari buruzko azterlanetako ebidentziak aztertuko ditut, bai eta, oro har, prozesu sentorial eta motorrei, ekintzari eta interakzio intersubjektiboari buruzko azterlanetakoak ere. José Bermúdezek arlo honetan egindako lana aipatuko dut, eta labur-labur azalduko dut zero puntuak irudikatu nahi zuena hobeto irudikatzen duen ordezkio nozio bat. Bermúdezen analisia baliagarri zaigu ohartzeko zero puntuaren kontzeptuari eustek ez duela zentzurik.

* **Correspondence to:** Shaun Gallagher. Department of Philosophy, Clement Hall 337, University of Memphis, Memphis, TN 38152. – s.gallagher@memphis.edu – <https://orcid.org/0000-0002-3147-9929>

How to cite: Gallagher, Shaun (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>).

Received: 12 May, 2024; Final version: 17 September, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

The concept of a “zero-point” (*Nullpunkt*) is pervasive in phenomenological accounts of perception (e.g., Husserl, 1989; Merleau-Ponty, 1964; Morley, 2001; Parnas *et al.*, 2005; Stein, 2012). The zero-point is meant to signify the bodily origin of the first-person perspective or the egocentric spatial frame of reference through which we encounter the world. Although it is meant to be a very embodied concept, since the zero-point is said to be located in or at the body, I’ll suggest that it is a complete abstraction that misleads the analysis of perception and action. Specifically, I’ll argue that body-schematic processes involved in perception and action are “thick,” multi-dimensional, or complex with varying perspectives running along axes relative to different parts of the body, and complicated by temporal and intermodal factors. Perception is better characterized as involving multiple trajectories rather than a zero-point.

I’ll start by indicating how the zero-point has been characterized in phenomenology, as well as in analytic philosophy of mind. I’ll consider several different characterizations: as an objective physiological point, as a geometrical point, or, phenomenologically, as part of one’s experience. I’ll introduce evidence from studies of vision, examine the concept of ‘zero-point aspects’ as outlined by Horgan and Nichols (2016) and then look more generally at sensory-motor processes, action, and intersubjective interaction. I’ll show that in all these cases a zero-point is not to be found. Finally, with reference to José Bermúdez’s work in this area, I’ll briefly outline an alternative notion that better captures what the zero-point was meant to capture. Bermúdez’s analysis helps us to see that there is no point to the concept of zero-point.

1. *Phenomenological accounts of the zero-point*

William James, in his *Essays in Radical Empiricism*, suggests that our experience of the world is accompanied “at all times with our body as its center, center of vision, center of action, center of interest. Where the body is is ‘here’; when the body acts is ‘now’; what the body touches is ‘this’; all other things are ‘theres’ and ‘thens’ and ‘thats’” (1976, p. 86n). This implies that the worldly things in one’s environment fit into a pervasive organization that refers back to an origin of perception, action, and interest located in the subject’s body, and that this orders all of one’s experiences, including thoughts and feelings, which “terminate in the activity of the body.... The body is the storm center, the origin of coordinates, the constant place of stress in all that experience-train” (1976, p. 86n).

Around the same time as James’s essays, Husserl tries to capture the same idea in the concept of the zero-point in his discussions of the lived body, and in a variety of contexts. For example, in his posthumously published *Ideas, Vol II* he writes:

[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)

Or again,

The lived body is, in the first place, the medium of all perception.... [O]bviously 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)

Accordingly, Husserl describes the zero-point as belonging to the lived body, and as involved in spatial perception, and specifically as providing the origin point of spatial orientation. He elaborates further about both our sense of orientation and direction in spatial experience, the zero-point’s relation to the body, and its connection to a persistent lawfulness in behavior.

[One's momentary focus] is necessarily related to the zero-point of orientation, to the absolute 'here', and to the concomitant system of the depth dimension (forward-back), and of the breadth and height dimensions (right-left, above-below). (1989, §32)

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'. It is thus that all things of the surrounding world possess an orientation to the body, just as, accordingly, all expressions of orientation imply this relation. (1989, §41)

It's clear that Husserl takes the zero-point to be an intrinsic feature of one's bodily and spatial experience, such that it has a phenomenological sense.¹

The notion of the zero-point is not a minor issue, since it seemingly characterizes the first-person perspective and is relevant not just to spatial experience, but to notions of the self (or ego), is referenced in analyses of psychiatric disorders (e.g., Henriksen *et al.*, 2019; Parnas *et al.*, 2005), and is also involved in characterizations of intersubjective empathy. As Edith Stein (2012, p. 63), for example, put it, "I experience the other as having another 'zero point of orientation' (*Null-Punkt der Orientierung*)."² One also finds the notion in Sartre: "[T]he perceptive field refers to a center objectively defined by that reference and located in the very field which is oriented around it. Only we do not see this center as the structure of the perceptive field considered; we are the center" (1956, p. 365). More generally it can be found as part of the phenomenological analyses that involve the character of memory or imagination insofar as these cognitive acts involve a kind of re-presentation (*Vergegenwärtigung*) of what has been or could be presented in perception (see, e.g., Morley, 2001).

A similar phenomenological account can be found in analytic philosophy of mind. I'll briefly point to two examples, the first one a very clear example that references Husserl's concept directly; the second, a more ambiguous example centered around Gareth Evan's work, and some issues that relate in some regard to developments in phenomenology, especially in the work of Merleau-Ponty (more of which later).

Consider first an important article by Terry Horgan and Shaun Nichols (2016), entitled "The zero point and I." I'll discuss this article in more detail later, but here let me just note that Horgan and Nichols point in two directions: back to Husserl, who they explicitly cite as the origin of this idea, and forward to a set of conceptual issues to which they think this idea has relevance. As they put it, "An enormously fruitful idea we will draw upon from the continental tradition is the one expressed by Husserl's expression 'der *Nullpunkt*' (the zero point)" (2016, p. 148). As they indicate, this is a fruitful idea precisely because it has relevance to a number of problems, including: the intrinsic aspects of phenomenality, phenomenal intentionality and non-reflective self-awareness; sensory-perceptual experience; voluntary-action experience – including self-agency/self as origin; the temporal structure of experience; and a variety of cognitive processes that involve desire, intention formation, and belief formation. I'll return to the issue of whether they remain faithful to Husserl's specific interpretation of the zero-point.

Another ongoing discussion of the zero-point in analytic philosophy of mind can be seen in Christopher Peacocke's (1992) notion of 'scenario content' which involves a point of origin, which he locates in the middle of the chest. Peacocke draws this idea from Gareth Evans's concept of the egocentric spatial frame of reference. There is, however, some debate in the literature about whether the egocentric spatial frame of reference actually involves a zero-point. On the one hand, Quassim Cassam (1997) explicitly links Evans' analysis of the egocentric spatial frame of reference to Husserl's zero-point and writes that "the Body ... is the zero point or point of origin of egocentric spatial perception" (1997,

¹ Note that Husserl uses the term 'zero-point' differently in a number of other contexts pertaining to affectivity and temporality (see, e.g., Husserl, 2012).

p. 53). Likewise, Lilian Alweiss (2018) recently made the same link defending Husserl's notion of the lived body as the zero-point in contrast to what she takes to be Evans and Cassam's reference to a material, objective body. On the other hand, Rick Grush (2005; see Briscoe & Grush, 2017) argues that Evans pre-figured sensory-motor contingency theories of the sort that one finds in Alva Noë (2004; also see Gallagher, 2024 for discussion of this point). As Evans puts it (1982, p. 160): "There is only one egocentric space, because there is only one behavioural space," emphasizing the correlation between perception and action and the experience of space. Thus, Evans states:

[... H]aving the perceptual information at least partly consists in being disposed to do various things ... The subject hears the sound as coming from such and such a position, but how is this position to be specified? We envisage specifications like this: he hears the sound up, or down, to the right or to the left, in front or behind, or over there. It is clear that these terms are egocentric terms: they involve the specification of the position of the sound in relation to the observer's own body. But these egocentric terms derive their meaning from their (complicated) connections with the actions of the subject ... (Evans, 1985, pp. 383-384)

I will not try to resolve the issue about how to interpret Evans, although I'll suggest that with Grush's interpretation we can start to see some complications that challenge the very notion of a zero-point. Some of these complications involve the fact that our mobile eyes are set within a head that can turn, and which is attached to a body that can turn, bend, assume different postures, and move from place to place. The concept of a zero-point tends to be modeled on a stationary point of view, the 'here' and 'now' which is only the limiting case of a mobile point of view. What we see, hear, touch, taste and smell is conditioned by what we do, and what we are capable of doing —our pragmatic possibilities based on the sensorimotor capacities of our bodies. "In ordinary experience, perception and movement are always united. I touch something by moving the arm. I see something by moving the head and eyes. What is perceived is perceived as nearby and perhaps reachable, or further way, as something that can be approached and explored" (Gallagher & Zahavi, 2020). Our perceptual organs (eyes, hands, ears, etc.) function in a way that is intertwined with our body's kinesthetic experience (Husserl, 2013). I'll argue that these are factors that complicate claims about a zero-point.

2. Three conceptions of the zero-point

It's well known that in western philosophical traditions vision is the dominant metaphor for cognition. That aside, the actual experience of vision suggests that there is something like an origination point; but what precisely is this point, and what do we mean by it? There seem to be three options.

- The zero-point is a real *physical* (physiological) point where visual information is processed
- The zero-point is a *geometrical* (ideal) point, the intersection of a set of lines that we might draw from objects back to our perceiving body
- The zero-point is a *phenomenological* point that we experience prereflectively as part of our visual experience

It will be useful to explore all of these options (also see section "Zero-points and self-location"). First, concerning a *physical zero-point*, consider a description of egocentric spatial organization given by Merker (2013, pp. 9-10). Attention is said to originate from a single point located inside our body, namely "at the proximal-most end of any line of sight or equivalent line of attentional focus." What we find in the objective physiology of normal binocular vision, however, is a real mix of processes happening across a complex set of anatomical structures. We might look for a physical zero-point in a number of these places. Might we find it at the optic nerve where retina is connected to the rest of the visual system? But, assuming we have two eyes, we operate two optic nerves, and a visual center cannot be in two places. The optic chiasm, then, where signals tend to meet? But the optic chiasm is a structure (not a zero-point) where a divi-

sion or diversification rather than integration occurs (signals from respectively left and right sides of the retinas crossing contralaterally rather than meeting up). Is there a zero-point in the cortex? It's well known that what we find in the visual cortex is distributed processing in different areas (V1, V2, V3 ...) and no one place (a Cartesian theater) where it all comes together. Moreover, if there were some place of visual binding, the binding itself would be of a complex visual scene and we would still have to specify some kind of zero-point related to that scene.

One possible candidate for a physiological zero-point related to vision —a mid-brain structure called the superior colliculus [SC] (in mammalian brains; the optic tectum in other vertebrates). Neurons in the surface layers of the SC receive direct input from retinas and almost exclusively respond to visual stimuli. SC is involved in behavioral responses toward specific points in egocentric space —the result of visual stimulation or attention. It contains one or more topographical maps correlated with retinotopic coordinates that are activated in ways that allow the organism to respond (e.g., with eye movements or head turns, or shifts of attention) to points in egocentric space (Katyal *et al.*, 2010).² The SC may be activated by visual input or by cortical activation.

Note, however, that the SC is bicameral —some neuroscientists refer to it in the plural; thus, “In each SC, the map represents the contralateral visual field” (Katyal *et al.*, 2010, 3075). This duality clearly challenges the idea that there is a zero-point located somewhere on the SC (or *colliculi*) maps. The fact that there is more than one map, and that the maps are organized in different layers complicates things. In the end, then, if we are looking for a clear physiological locus for the zero-point, it's not clear where we can find it. This lack of a physical candidate also highlights the fact that if there is an experiential zero-point (visually or otherwise) it is not necessarily the case that its neurophysiological correlate (if there is such a thing) will be isomorphically a zero-point.

When we start to think of different types of experiences —not just perception and its spatial framework— the notion of a localizable zero-point gets more obscure. As Marchetti (2024, p. 8) points out, citing Revonsuo (2006), “Our emotions are located somewhere in our body; we feel our memories as originating and located in ourselves; we have ideas and thoughts in our mind, etc.” But it's difficult to point to any origination points for any of these experiences. More generally, the zero-point remains “elusive” (Williford *et al.*, 2012; 2018; Marchetti, 2024). If the zero-point were a physical thing, like an eye, even then it “is excluded from the contents of consciousness by the same geometric necessity that prevents an eye from viewing itself, though it is the instrument for viewing all else” (Merker, 2013, p. 10).

The idea of a *geometrical zero-point* has been proposed by Blanke and Metzinger (2009) to characterize the origin of the seeing subject's embodied perspective. They call this a weak first-person perspective which could be understood as “a purely geometrical feature” of our visuospatial presentation of reality. This would be an ideal (unreal) geometrical point, located on some virtual or ideal axis. Horgan and Nichols (2016) refer to it as “Husserl's mathematical metaphor” and they link it to Cartesian coordinates, i.e., where the x and y axes meet. But this does not seem to be what Husserl means by zero-point since his claim is that the “lived body has the unique distinction of bearing in itself the zero point (*Nullpunkt*) of all these orientations” (1989, §41a). Accordingly, Husserl's claim is not purely formal or geometrical; it's about something that we literally experience —an *experiential zero-point*. Merleau-Ponty makes this clear: for phenomenology,

Space is no longer what it was in [Descartes'] Dioptric, a network of relations between objects such as would be seen by a witness to my vision or by a geometer looking over it and reconstructing it from outside. It is, rather, a space reckoned starting from me as the zero point or degree zero of spatiality. I do not see it according to its exterior envelope; I live in it from the inside; I am immersed in it. After all, the world is all around me, not in front of me. (Merleau-Ponty, 1964, p. 178)

² I thank Sucharit Katyal for calling this to my attention at a conference on the First-Person Science of Consciousness – Theories, Methods, Applications. Witten, Germany (23-25 May 2019).

The idea that the zero-point is part of our experience is also defended by Barry Dainton who proposes that “in any perceiving, experiential contents of whatever type can only seem to be presented to a subject if the subject itself has the impression of being itself spatially related to what it is perceiving. Ordinary human phenomenal fields are thus centred rather than centreless” (Dainton, 2016, p. 130). More than perception is included in human phenomenal fields, however. There are obviously other experiences that are also characterized by first-personal phenomenality. Consider feelings of joy, despair, or shame. Does it seem that their phenomenality is more centered than distributed?

What could count as an embodied experiential zero-point? There are several candidates that we can consider. For example, could the zero-point correspond with the blind spot? It seems not, since that point is both outside the body and outside of our phenomenology, that is, we don’t experience it in normal vision. Could it be at the experienced focal point of our attention? This would seem to be on the wrong side of the intentional relation. That is, Husserl describes the zero-point as something pertaining to the subjective noetic aspect of consciousness (the focusing) rather than to the noematic side, that is, the intentional (focused-upon) object in the perceptual field. 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.³ 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? Philosophers like Dorotheé Legrand (2011) and Helena De Preester (2007) contend that this prereflective and non-transitive awareness of one’s own body is not of a bare orientation point but of an experienced volume, a kind of substantial thickness of one’s “deep” body. In this regard, the experiencing subject is more than a perspectival zero-point; phenomenologically, it is something > 0 .

Also, vision is never just bare vision; vision always involves cross-modal connections with other senses. Husserl (2013) indicates that it involves kinaesthesia. For example, ocular muscles operate such that they allow us to focus, providing an implicit sense of movement that is part of visual experience. Perception more generally is intermodal —vision, for example, registers in other sense modalities, in the same way that taste is not strictly on the tongue. Hearing is in stereo since we typically listen with two ears, and it’s the position and distance between our ears that define our sense of where the source of the sound is. Likewise, if we are moving or if the object is moving, we might think that a trajectory rather than a zero point is involved.

The idea of a trajectory is related to the fact that there is an underlayment to perception that includes bodily dimensions and motoric processes. Vision is embedded in an articulated hierarchy of moving parts and degrees of freedom: the eyes move; the head turns; the torso turns; I can move up or down or laterally to change my perspective. Vision is thick with both the possibility and actuality of motoric kinaesthesia. This cross-modality is also reflected in the organization of the superior colliculus in which deeper neuronal layers involve multi-sensory integration.

3. *The sensorium moves, and a clue from Bermúdez*

I suggest that having *two* ears is like having two eyes and, accordingly, the auditory sensorium also operates with complex disparities. Taste is no simpler; there are (at least) *three* areas of the palate and tongue (as wine connoisseurs know well), with taste also a distributed process that involves at least *one* nose. It’s difficult to find any 0s in the accounts of any of these senses.

³ 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.

In touch there are multiple lines of orientation since touch potentially involves not only two haptically oriented hands but can also involve different parts of the body as well as proprioception and kinaesthesia. This becomes clear when we consider haptic movement.

Indeed, one might suspect that the zero-point as a standard of orientation is better conceived in motoric terms, or in terms of enactive sensory-motor contingencies. In this regard, Merleau-Ponty weighs in. Like Gareth Evans (at least on one interpretation), he is not in favor of the zero-point. Merleau-Ponty discusses, but never takes up Husserl's notion of the zero-point of orientation. In his Sorbonne lectures (2010), in a running commentary on Husserl, he summarizes Husserl's idea of the zero-point without endorsing it (2010, p. 75); he later uses it to describe von Uexküll's view of the *Umwelt*. Rather than the zero-point Merleau-Ponty discusses a more holistic body schema as the organizer of the perceptual field; the body enacts a "motor-perceiving" (2010, p. 216). There is, first,

the postural schema. All the elements of position totalized are either in series or simultaneously... and referred to the space of the world. Second, more generally there is an intersensorial system of equivalences that functions as a whole.... Third, there is an account taken of my movements.... The corporal schema furnishes me with the tracing that intervenes in the reading of the world (e.g., the active movements of the eye and the rest of the world despite the movement of images). (2010, p. 217)

If perception is enactive, even if it is visual perception, it involves not just eyes but bodily position, hands and the degrees of freedom that define whether I can reach the object or not. What allows for this is a thick and articulated schema, rather than a zero-point. In the *Visible and Invisible*, Merleau-Ponty talks about the body as "a dimensional this" —a "field" (1968, p. 260). He takes Husserl's well-known example of one hand touching the other hand and then reversing the touching and touched (what Merleau-Ponty calls 'reversibility') and he finds a shifting difference (*écart*) between them rather than a zero-point. The touching never hits zero —it is always on the verge, always "offset" — he calls it the "zero of being which is *not nothingness* (p. 260) —a zero that is always something more than zero.

Merleau-Ponty characterizes movement in a way that undercuts anything like a zero-point:

The movements of one's own body are naturally invested with a certain perceptual signification, they form a *[well-articulated] system with external phenomena* so tightly woven that external perception "takes account" of the movements of the perceptual organs, and it finds in them, if not the explicit explanation, then at least the motive for the intervening changes in the spectacle and can thereby understand these changes immediately (Merleau-Ponty, 2012, p. 49; emphasis added)

Charles Taylor draws on Merleau-Ponty and explains this perhaps more clearly:

Our perceptual field has an *orientational structure* . . . In those rare moments where we lose orientation, we don't know where we are; and we don't know where or what things are either; we lose the thread of the world, and our perceptual field is no longer our access to the world, but rather the confused debris into which our normal grasp on things crumbles. . . . (Taylor, 1995, p. 23; emphasis added)

This kind of confusion seems unlikely if we carry around a zero-point that anchors our perspective.

It is not just that the field's perspective centers on where I am bodily —this by itself doesn't show that I am essentially an agent. But take the up-down directionality of the field. What is it based on? Up and down are not simply related to my body; up is not just where my head is and down where my feet are. For I can be lying down, or bending over, or upside down . . . I have to maintain myself upright to act, or in some way align my posture with gravity. Without a sense of 'which way is up', I falter into confusion. (Taylor, 1995, p. 23)

I note that in contrast to up and down, left and right may be more firmly anchored to one's body, not just in regard to what Kant pointed out about hands being incongruent counterparts (Kant, 1992; Gallagher, 2006), but in terms of the right always being on one's right (to the right of one's body) and the left being on one's left, and this doesn't matter if we turn around, or are upside down. If we are more anchored by our body in these lateral directions, that fact can only establish a centered axis. We would need to have a similar arrangement with up and down (which we don't, as Taylor indicates) to establish something like the reticle or crosshair of a zero-point.

José Bermúdez (2005) also gives us reason to move away from the idea of a single zero-point, or single egocentric spatial frame of reference. He suggests three frames of reference.

1. **The object-relative spatial frame of reference.** For example, when reaching one's hand to grasp an object there is a trajectory from the initial position of the hand to the position of the relevant object. That trajectory links hand and object in the same "object relative" spatial framework, which is equivalent (following Evans) to what is standardly called the *ego- [or body-] centric spatial frame of reference*.

The important point is that in the object-relative spatial frame of reference the orientation is centered on the hand that is reaching. If the other hand was doing the reaching, there would be a different hand-centric frame of reference. Likewise in the case of kicking with my left versus right foot, etc.

In addition, however, Bermúdez identifies two more spatial frames of reference.

2. **The internal spatial framework:** in some kinds of actions, one has to calibrate between the object/body-centric framework and an internal organization (keeping track of one limb's relative to other limbs). His example is returning a volleyball—we need to keep track of the ball in egocentric space, but also we need to keep track of how our hands relate to one another.
3. **The orientation frame of reference.** This is equivalent to what Charles Taylor describes in reference to gravity and orientation (involving vision, kinaesthesia, and vestibular processes for position and balance).

These three spatial frameworks need to be integrated in our experienced action. Accordingly, we get a three-fold, more complex gestalt structure rather than anything as simple as a zero-point. A good example of this is something we do everyday: dressing. Putting on your clothes involves a complex set of movements that require coordination in all three spatial frames of reference.

Even if we focus on just the first, object-relative or egocentric frame of reference, there is a complication. When I reach to grab an object I usually do so as I visually focus on it, or at least make a quick glance to determine its location. So, simultaneously it seems I have a visual zero-point seeming centered in the neighborhood of my eyes, plus a zero-point centered at my hand. Imagine the act of punting in a game of American football where you are preparing to catch and then kick the ball that is coming toward you—there would be three object-relative zero-points involving eyes, hands, and foot—all coordinated by an articulated and dynamically changing body schema rather than any one stationary zero-point. Mathematically, of course, $0(n) = 0$; or $0+0+0 = 0$. But that's not the rule here. A plurality of zero-points is not equivalent to a zero-point; $0+0+0 =$ a bodily configuration.

The above-zero complexity increases when we fold in the other spatial frameworks identified by Bermúdez. When I reach to grab an object, as he indicates,

it is natural to suggest that the axes that determine particular proprioceptive frames of reference are centered on particular body-parts, just as are the axes determining the frames of reference for perceptual content and basic intentions. The picture that emerges, therefore, is of a number of different representations of space, within each of which we find representations both of bodily and of non-bodily location. So, for example, we might imagine reaching behavior to be controlled

by an egocentric frame of reference centered at some location on the hand.... Despite its appealing economy, however, this account is ultimately unacceptable, because of a fundamental disanalogy between the bodily space of proprioception and the egocentric space of perception and action. (2005, p. 309)

Proprioceptive awareness of our own body is not egocentric —it's not oriented around an origin. Even haptic touch reflects this lack of a zero-point. Consider that I can attend to an environmental object with my haptic hands. With my hands I can shift attention from the tactile feel of the object to the temperature I experience on my skin, or from what the object feels like, to what my hand feels like as I move it on the object. In this case there is no one zero-point that organizes my exploration; nor is there any shift from one clearly defined point to another. If in our visual attention we were to shift from point A in the visual field, to point B in the visual field, one might think that a zero-point is formed at the vortex located near the eyes. If this is a zero-point, it's not the type of zero-point that could be associated with my hands' haptic touch. My hands, moreover, are involved in both my tactile and temperature attending experience that may be distributed across several fingers or several parts of my hand – they are not confined to one point, and they are not oriented around one point in anything like a vortex arrangement. Rather, as my body, my eyes, my hands move throughout this exploration it's a matter of a complex set of trajectories rather than any one zero-point.

What Merleau-Ponty calls reversibility, or the enveloping relations that characterize my body (1968, p. 260), disrupts any concept of zero-point. The kind of doubling of sensation in hand-touching-hand is not zero-pointed, but ambiguously distributed. In this case, attention does not originate in my head or chest; attention is in my hands; and my hands are not located in an egocentric spatial framework —the lived-proprioceptive location of my hands belongs to a global proprioceptive spatial framework (Bermúdez, 2005; Gallagher, 2006). According to Merleau-Ponty, “the spatiality of the body must descend from the whole to the parts, my left hand and its position must be implicated in an overall bodily plan and must have their origin there” (2012, p. 101). The spatiality of the body “is not, like the spatiality of external objects or of ‘spatial sensations’, a positional spatiality; rather, it is a situational spatiality” (2012, p. 102).

One can say, as Merleau-Ponty does, that perception is indexed to the body as a center. But given how bodily space is organized, this is different from saying that it is indexed to a zero-point. A zero-point defined in objective, Cartesian or geometrical terms, cannot be found in the perceiving or acting body given the complexities mentioned above. As Bermúdez points out, the proprioceptive organization of the body is not equivalent to an objective geometry that one might associate with the egocentric frame of reference (1998, pp. 152-153). If, as he indicates, proprioception does not organize the differential spatial order of the body around an origin, this means that whereas one can say that this book is closer to me (my body) than that book over there, one cannot say that my foot is closer to me than my hand. As Merleau-Ponty puts it, “When the word ‘here’ is applied to my body it does not designate a determinate position in relation to other positions or in relation to external coordinates ...” it expresses “the situation of the body confronted with its tasks” (2012, pp. 102-103).

4. *Zero-points and self-location*

Despite all of these complexities involved in the sensory-motor aspects of perception and action it may still seem highly intuitive that our spatial experience is organized around a zero-point that we prereflectively experience as such. Indeed, there have been empirical and microphenomenological studies that claim to show that there is something like a zero-point, or a precise location of the experiencing self. Consider, for example, the study by Franco Bertossa *et al.* (2008). They used what is now called a microphenomenological, structured interview method to investigate where subjects would locate their perceiving I (the “I-that-perceives”). It turns out that a high percentage (83%) of subjects located their perceiving I in the temporal area of the head centered midway behind the eyes. Subjects included fifty-four (54) sighted western individuals, eight (8) blind subjects and five (5) non-western subjects. This

looks like strong evidence that there is a phenomenological sense of a specific perceptual zero-point – precisely the kind of thing that Husserl was describing. Some of the details of the experiment, however, should lead us to discount this idea.

The interview procedure starts by asking about the location of objects in egocentric space, specifically asking about the distance of objects from one's body. The interview then probes the subject's sense of up and down, and then their sense of themselves as a perceiving subject. During the interview subjects are tested on their proprioceptive ability (with their eyes closed) to "recognize" various parts of their body. Just at this point, however, there is a subtle shift in the way the interviewer refers to the subject. The interviewer had been using the pronoun 'you'. For example, the interviewer asks: "Can you tell me where the ceiling is with respect to you?" Or, "How close is this object to you?" The shift then occurs when the interviewer begins to reference "the I." Indeed, it is an odd question that the interviewer asks: "Is your hand 'I', or does your 'I' perceive through your hand? Although the scare quotes around the 'I' appear in the textual report, it's not clear what they signify, or how those scare quotes were handled during the interview. Using the idiom of 'the I' the interviewer then asks about the location of various body parts relative to the 'I'. For example, the interviewer asks: "Is there a zone between your belly and your throat that you feel is closer [to the I]?" Subjects then say that the throat is closer. Pursuing this line of questioning the subject is led to actually point to a location of the 'I' inside the head.

Although it is not absolutely clear in Bertossa *et al.* (2008), it seems that during the interview the I becomes reified and subjects are then asked to locate various body parts relative to this reified I, and finally to locate the I itself. One has to wonder whether the shift to talking about 'the I' isn't similar to the shift made by Descartes between the *cogito* and the '*res cogitans*'. What would happen if the interviewer stayed with the pronoun 'you'? The question would be something like the following: "Is your ankle closer to you than your chin?" It is quite possible to answer that question by saying "My ankle and chin are both part of me; one is not closer to me than the other." Indeed, at least one subject tried to answer in this way by saying that "the I" corresponded to his whole body, at which point the interviewer asks him for an explanation: "why do you say that your 'I' coincides with your whole body?" to which the subject answers "I really have no idea." (p. 332). It's not clear, however, that this last 'I' (who has no idea) is the same 'I' that is being located.

In other words, it is not clear that such a reified 'I' is what Husserl discusses when he associates the lived body with the zero-point. I suggest that something about the prereflective nature of experience gets lost in the procedure of trying to locate the zero-point. Bertossa *et al.*, however, have no doubt that the 'I' is "placed somewhere near the centre of [the] head. This seat corresponds to a 'point zero', the origin (0,0,0) point of a Cartesian spatial geometric framework, whose axes are defined by the subjects' experience of what lies closer or farther in front or behind, above, or below, where their sense of 'I' and their sense of 'here' are felt to coincide" (2008, p. 333).

Alsmith and Longo (2013, p. 71) note that "the inherent diversity and flexibility of multi-sensory integration processes" can lead subjects to locate the self in different locations in the body. They also note that the study by Bertossa *et al.* (2008) doesn't allow for this possibility. In their own study they asked subjects to adjust a pointer to indicate the experienced location of the self.

So which part of the body counts as you? Our results suggest that no single body part is judged as the unique seat of the self. Beyond a certain spatial resolution, self-location judgements are ambiguous between at least two locations, though when forced to judge this ambiguity can be resolved according to contextual factors. When asked to point directly at themselves, in both haptic and visual modalities, our participants' judgements were clearly affected by the starting location of the pointer. Participants most frequently chose to point to one of two likely parts of the body according to which they reached first, upper face or upper torso.... To the extent that such self-location judgements tell us anything about our concept of the self as a spatial entity, they tell us that the concept is inherently ambiguous. (Alsmith & Longo, 2013, pp. 73-74)

There are other studies that attempt to specify the location of the self. The self, of course, can be defined in numerous ways and in such studies the goal is not necessarily to find a zero-point-self. In some ingenious whole-body (or out-of-body) illusion experiments conducted by Olaf Blanke's group (e.g., Ionta *et al.*, 2011), experimenters were able to manipulate the location of the first-person perspective and self-location where self is equated with the whole body. In their questionnaires they ask questions such as "How strong was the feeling that the body you saw was you?" In this respect, although they are asking about bodily location, they are not reifying the self. They found that activation in the left and right Temporal Parietal Junction (TPJ) correlated with changes in experienced changes in body location. In such whole-body illusion experiments where, for example, the subject feels as if they are floating, the first-person perspective follows the experienced body. One modulates the first-person perspective and self-location by manipulating tactile, visual, somatosensory and vestibular contributories, which implies that the first-person perspective is not simple, but is underpinned by complex processes. As Ionta *et al.* indicate, the TPJ integrates multisensory signals, and its "receptive fields are most often large and bilateral, may encompass the face, trunk, hemibody, or entire body, and have bimodal visuo-tactile receptive fields that are anchored to the body" (2011, p. 371). But the TPJ, like the SC, is just one of a large number of brain areas activated for self-related phenomena (see Legrand & Ruby, 2009; Vogeley & Gallagher, 2011). It may be closer to the truth to say that in whatever way we define the self it involves at least the whole brain together with the whole body (see Gallagher, 2013).

5. *Attending or acting with others*

Before I suggest a better way to deal with the complexity and ambiguity of the perceiving and acting self that the notion of a zero-point is supposed to fix, let me add one more complicated dimension. If one's individual perception or action has a complex gestalt-like framing, this can be even more complex in cases of joint attention and joint action. We can think of joint attention as a "basic joint action" (Fiebig & Gallagher, 2013), and then consider that in most joint actions one needs to coordinate one's movements with others. For example, as two people engage in moving a bookcase through a doorframe all three spatial frames of reference identified by Bermúdez, *times two*, are involved. The coordination of my own two-handed grip, together with my general balance, together with my perception of the doorframe through which we need to navigate, and whatever ongoing movements we might have to make to adjust the angle of the bookcase to ease it through, must be coordinated with your movements which in some instances may throw me off balance. In joint actions that involve two-person synchronic, correlated movement patterns, one's lived body schema meshes with the other person's to form a "joint body schema" —which involves an integration of one's peripersonal space with the other's (Soliman & Glenberg, 2014). This is an explicit case of what Merleau-Ponty calls intercorporeity, and it involves a complex gestalt of 3x2 spatial frames of reference rather than two individual or one collective zero-point(s).

Similar things can be said with respect to communicative interactions. Indeed, Jürgen Habermas makes a related suggestion, describing an organization of space surrounding two or more speakers. He suggests that it is the speech situation, encompassing two or more speakers, rather than the body of the individual subject, that is the center around which the context is ordered. We could think of this in terms of Merleau-Ponty's concept of "situational spatiality." In such situations, the first-person standpoint is really taken up into a second-person standpoint, in both an experiential and spatial sense, as we are speakers and actors *alongside others*.

The common speech situation constitutes the center —and not, for instance, my body, as an anthropologizing phenomenology has claimed— in which social spaces (staggered concentrically according to depth and width) and historical times (arranged three-dimensionally) converge prior to any objectivation through measuring operations.... I, in my body, and I, as my body, find myself always already occupying an intersubjectively shared world, whereby these collectively inhabited lifeworlds telescope into each other, overlap, and entwine like text and context. (Habermas, 1998, p. 244)

This “always already” situation of intersubjectivity complicates the concept of a zero-point and precisely points away from the very notion.⁴

To summarize, the idea of an experiential zero-point characterizing embodied spatial experience is not to be found in the sensory realm (vision, touch, etc.), in the motoric realm of action, in the realm of self-experience, or in the social realm of interaction, joint action, joint attention, or communicative interaction.

6. *The alternative: The agentive situation*

Let’s consider one more attempt to make sense of the notion of a zero-point, and then turn to an alternative proposal. Horgan and Nichols (2016) attempt to stay with the notion of zero-point (drawing from Husserl), but they almost immediately start to use the phrase ‘zero-point aspects’. For example, under this heading they include two of Bermúdez’s three spatial frameworks, allowing for (1) a plurality of zero-points (e.g., associated with hands in haptic exploration, and (2) the orientation framework (*à la* Taylor and Bermúdez).

For instance, the up/down axis of one’s visual-presentational experience seems heavily dependent upon certain aspects of tactile and kinesthetic sensory-perceptual experience: roughly, down is the direction of the surface to which one’s body at rest is attached, and toward which one’s body moves when unattached to any surface. This up/down axis [is involved in perception of] objects in the ambient environment as located within an implicit reference frame that includes the directions of up and down. The up/down axis also is a zero-point aspect of visual-perceptual experience. (Horgan & Nichols, 2016, pp. 149)

Recall, as noted by Taylor, the up-down axis involves not just bodily directionality, but also environmental constraints, such as gravity. Whatever this axis is, it experientially extends into the environment. Horgan and Nichols call this a zero-point aspect. They continue:

Some aspects of external-world-representing tactile experience seem *centered on one or several zero points* that coincide with certain specific body parts that themselves are also explicitly represented tactilely; for instance my tactile experience of the dumbbell’s shape seems centered on my two hands, one of which feels (via gripping) the cylindrical shaft and the other of which feels the shape of one of the weight-plates to which the shaft is attached. Other aspects of external-world-representing tactile experience —for instance, my tactile experience of the dumbbell’s heaviness, and my tactile experience of the balance of this heaviness on its two ends— seems centered upon *a zero point that coincides with my entire body* (Horgan & Nichols, 2016, p. 149; emphasis added)

Although they suggest that “[o]ne experiences being located at a specific sensory-perceptual zero point ...” —it doesn’t seem that a zero-point aspect is a zero-point if it’s a *linear axis* or a trajectory in movement, or if it coincides with the *whole body*. In their insightful analysis of a number of phenomena, Horgan and Nichols define the notion of ‘zero-point aspects’ to signify those pre- or non-reflective aspects of self-experience that do not involve explicit self (I*) representation —but it is not clear what the concept ‘zero-point’ *per se* adds to their analysis, and it does not seem equivalent to their starting point in Husserl’s use of the term. Accordingly, I suggest that there is zero point to using the term ‘zero-point’ in such cases.

⁴ In contrast, Schutz and Luckmann borrow Husserl’s notion of the zero-point, the center of a system of spatial coordinates, use it to define what they term the ‘zone of operation’ (similar to George Herbert Mead’s notion of ‘manipulatory area’), and apply these concepts in social analysis (Schutz & Luckmann, 1973, pp. 36, 41ff).

How then should we think of these specific aspects of spatial experience in perception and action? Should we take the various objections against the zero-point to be a mere terminological issue? I think it is more than a terminological issue. The point of questioning the zero-point concerns the role of the first-person embodied perspective and its phenomenology. The first-person perspective seems not to be reducible to a zero-point, and the lived body is more than a collection of zero-points; more also than a geometrical *res extensa*. It involves an experiential “thickness,” an ambiguity, as well as temporally attuned motor-control processes and situated affective processes that enable and delimit perception, and action. In contrast to an abstract origin-point anchoring the “thin” lines of a geometrical mapping of directional attention in perception or action, experience extends across a multitude of intersecting dimensions that include affective and interoceptive processes (De Preester, 2007; Tsakiris & De Preester, 2018), as well as multiple affordances that configure the agent’s situation (for example, what the agent can reach and grasp while keeping an eye on what she can’t yet reach and at the same time attending to what her partner can or can’t reach and grasp). Such aspects get lost in an analysis that focuses on the abstract concept of zero-point. This abstraction is reflected in the fact that, in the end, Horgan and Nichols revert to discussions of abstract cognitive architecture rather than embodied subjectivity to explain zero-point perspective – on their view such perspective is “cognitive-architecturally implemented by zero-point representations.”

What such an analysis and the notion of a zero-point itself overlook is what I’ll call, drawing on Merleau-Ponty’s notion of a “situational spatiality,” the notion of an *agentive situation*.⁵ 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). Consider the example of initiating a tennis serve. In setting up the serve I may be marginally aware that I’m swaying back and forth on my feet which habitually transitions into a posture where the action of my throwing hand moves in close synchrony with the swing of my racket which is gripped in my other hand. My visual attention is on a specific zone located on the other side of the net while at the same time I’m aware of my opponent’s position and am anticipating my subsequent response to his return. Some interoceptive processes may be distracting me if my stomach is upset, or I have a painful injury that requires that I adjust my movements. Where should we locate the zero-point in this situation, which spans subpersonal processes of motor intentionality measured in elementary time scales of msecs, integrative processes of present conscious intentions-in-action measured in seconds, as well as longer-term distal intentions that define it in terms of playing tennis to win? Even if we considered a postural snapshot of some moment of the action, unless we equate the zero-point with the agent’s abstract center of gravity at that time-point, there is no one point on which to center the action; there is at the very least a complex set of coordinations —one foot in relation to the other, two hands in relation to each other while being in relation to ball and racket, respectively, while the agent is perceptually attending to a targeted zone and the other player. 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.

This idea of the agentive situation is partially captured in three statements that we have already seen, from Merleau-Ponty, Taylor, and Habermas, respectively.

The movements of one’s own body are naturally invested with a certain perceptual signification, they form *with the external phenomena* a [well-articulated] system. (Merleau-Ponty 2012, 49; emphasis added)

⁵ For a fuller explanation of this concept, drawing from John Dewey’s concept of situation, see Gallagher (2020, 1.2).

Our perceptual field has an *orientational structure* . . . It is not just that the field's perspective centers on where I am bodily —this by itself doesn't show that I am essentially an agent.... I have to maintain myself upright to act, or in some way align my posture with gravity. (Taylor 1978, 23)

I, as my body, find myself always already occupying an intersubjectively *shared world*, whereby these collectively inhabited lifeworlds telescope into each other, overlap, and entwine like text and context" (Habermas, 1998, p. 244)

The intertwinement of multiple dimensions —cognitive, affective, motoric, perceptual, interoceptive, to name these factors in the most general terms— constitutes a situation that has no synchronic center, but is rather a matter of dynamical trajectories that intersect as a set of processes that involve zero zero-points. This is not to deny that there is a subject of perception or an agent of action, or that they are one and the same, experientially located in “active coordinates,” “in virtue of the mutually structuring interrelation between his perceptual experience and his basic purposive interaction with the environment” (Brewer 1992, pp. 30-31).⁶

Acknowledgements

The author's research for this paper was supported by a Humboldt Foundation award and funding received as a Visiting Professor, Psychology, University of Rome-Sapienza.

REFERENCES

- Alsmith, A. J. & Longo, M. R. (2014). Where exactly am I? Self-location judgements distribute between head and torso. *Consciousness and Cognition*, 24, 70-74.
- Alweiss, L. (2018). Embodiment and self-awareness: Evans, Cassam and Husserl. *Philosophy*, 93(1), 31-51.
- Bermúdez, J. L. (2005). The phenomenology of bodily awareness. In D. W. Smith and A. Thomasson (Eds.), *Phenomenology and philosophy of mind* (pp. 295-322). Oxford University Press.
- Bermúdez, J. L. (1998). *The paradox of self-consciousness*. MIT Press.
- Bertossa, F., Besa, M., Ferrari, R. & Ferri, F. (2008). Point zero: A phenomenological inquiry into the seat of consciousness. *Perceptual and Motor Skills*, 107(2), 323-335.
- Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, 13(1), 7-13.
- Brewer, B. (1992). Self-location and agency. *Mind*, 101(401), 17-34.
- Briscoe, R. & Grush, R. (2017). Action-based theories of perception. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2017/entries/action-perception/>.
- Cassam, Q. (1997). *Self and world*. Oxford University Press.

⁶ In this respect, I think the enactive view that I am defending in this paper is very much in line with Bill Brewer's argument that agency is the ground of self-location: “perceptual experience carries self-locating spatial contents in virtue of its role in controlling and coordinating [the agent's] purposive interaction with the perceived environment, where two crucial aspects of this role are a reflection of the continuous dependence of the nature of experience on the changing position of the subject with respect to its objects and a sensitivity to the relevant biomechanical properties of the body in the production of appropriate spatial behavior (1992, 32).

- De Preester, H. (2007). The deep bodily origins of the subjective perspective: Models and their problems. *Consciousness and Cognition*, 16(3), 604-618.
- Evans, G. (1982). *The varieties of reference*. Oxford University Press.
- Evans, G. (1985). *Collected papers*. Oxford University Press.
- Fiebach, A. & Gallagher, S. (2013). Joint attention in joint action. *Philosophical Psychology*, 26(4), 571-87.
- Gallagher, S. (2024). Molyneux and motor plasticity. *Philosophy and the Mind Sciences* 5. <https://doi.org/10.33735/phimisci.2024.11596>
- Gallagher, S. (2020). *Action and interaction*. Oxford University Press.
- Gallagher, S. (2013). A pattern theory of self. *Frontiers in Human Neuroscience*, 7(443): 1-7. doi: 10.3389/fn-hum.2013.00443.
- Gallagher, S. (2006). The intrinsic spatial frame of reference. In H. Dreyfus & M. Wrathall (Eds.), *The Blackwell companion to phenomenology and existentialism* (pp. 346-355). Blackwell.
- Gallagher, S. & Zahavi, D. (2020). Phenomenological approaches to self-consciousness. In Edward N. Zalta (ed.), *Stanford encyclopedia of philosophy*. <https://plato.stanford.edu/entries/self-consciousness-phenomenological/>.
- Grush, R. (2005). Brain time and phenomenological time. In A. Brook & K. Akins (eds.) *Cognition and the brain: The philosophy and neuroscience movement* (pp. 160-207). Cambridge University Press.
- Habermas, J. (1998). *On the pragmatics of communication*. Maeve Cooke (Ed). MIT Press.
- Henriksen, M. G., Parnas, J. & Zahavi, D. (2019). Thought insertion and disturbed for-me-ness (minimal selfhood) in schizophrenia. *Consciousness and Cognition*, 74(102770), 1-9. doi.org/10.1016/j.concog.2019.102770
- Horgan, T. & Nichols, S. (2016). The zero point and I. In S. Miguens, G. Preyer & C. Bravo Morando (Eds.), *Pre-reflective consciousness: Sartre and contemporary philosophy of mind* (pp. 143-175). Routledge.
- 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.
- Husserl, E. (2012). *Analyses concerning passive and active synthesis: Lectures on transcendental logic*. (A. Steinbock, Trans.). Springer.
- Husserl, E. (2013). *Thing and space: Lectures of 1907*. (R. Rojcewicz, Trans.). Springer.
- Ionta, S., Heydrich, L., Lenggenhager, B., Mouthon, M., Fornari, E., Chapuis, D., Gassert, R. & Blanke, O. (2011). Multisensory mechanisms in temporo-parietal cortex support self-location and first-person perspective. *Neuron*, 70(2), 363-374.
- James, W. (1976). *Essays in radical empiricism*. Harvard University Press.
- Kant, I. (1992). Concerning the ultimate ground of the differentiation of directions in space. In D. Walford & R. Meerbote (Eds.), *The Cambridge edition of the works of Immanuel Kant. Theoretical philosophy, 1755-1770* (pp. 365-72). Cambridge University Press.
- Katyal, S., Zughni, S., Greene, C., & Ress, D. (2010). Topography of covert visual attention in human superior colliculus. *Journal of Neurophysiology*, 104 (6), 3074-3083.

- Legrand, D. (2011). Phenomenological dimensions of bodily self-consciousness. In S. Gallagher (Ed.), *Oxford handbook of the self* (pp. 204-227). Oxford University Press.
- Legrand, D. & Ruby, P. 2009. What is self-specific? Theoretical investigation and critical review of neuroimaging results. *Psychological Review*, 116(1), 252-82.
- Marchetti, G. (2024). The self and conscious experience. *Frontiers in Psychology*, 15:1340943. doi: 10.3389/fpsyg.2024.1340943
- Merker, B. (2013). The efference cascade, consciousness, and its self: Naturalizing the first person pivot of action control. *Frontiers in Psychology*, 4, 501.
- Merleau-Ponty, M. (2012). *Phenomenology of perception*. (D. A. Landes, Trans.). Routledge.
- Merleau-Ponty, M. (2010). *Child psychology and pedagogy: The Sorbonne lectures 1949-1952*. (T. Welsh, Trans.). Northwestern University Press.
- Merleau-Ponty, M. (1968). *Visible and invisible*. (A. Lingis, Trans.). Northwestern University Press.
- Merleau-Ponty, M. (1964). *The primacy of perception*. (J. Edie, Trans.). Evanston, USA: Northwestern University Press.
- Morley, J. (2001). Inspiration and expiration: Yoga practice through Merleau-Ponty's phenomenology of the body. *Philosophy East and West*, 51(1), 73-82
- Noë, A. (2004). *Action in perception*. MIT Press.
- Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179-217.
- Parnas, J., Møller, P., Kircher, T., Thalbitzer, J., Jansson, L., Handest, P. & Zahavi, D. (2005). EASE: examination of anomalous self-experience. *Psychopathology*, 38(5), 236-258.
- Peacocke, C. (1992). Scenarios, concepts, and perception. In Tim Crane (Ed.), *The contents of experience* (pp. 105-135). Cambridge University Press.
- Revonsuo, A. (2006). *Inner presence. Consciousness as a biological phenomenon*. MIT Press.
- Sartre, J-P. (1956). *Being and nothingness: An essay on phenomenological ontology*. (H. E. Barnes, Trans.). Philosophical Library.
- Schutz, A. & Luckmann, T. (1973). *The structures of the lifeworld*. Northwestern University Press.
- Soliman, T. & Glenberg, A. (2014). The embodiment of culture. In L. A. Shapiro & S. Spaulding (Eds.). *The Routledge handbook of embodied cognition* (pp. 207-220). Routledge.
- Stein, E. (2012). *On the problem of empathy*. Springer.
- Taylor, C. (1995). *Philosophical arguments*. Harvard University Press.
- Vogeley, K. & Gallagher, S. (2011). The self in the brain. In S. Gallagher (ed.), *The Oxford handbook of the self* (pp. 111-36). Oxford University Press.
- Williford, K., Bennequin, D., Friston, K. & Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Front. Psychol.* 9:2571. doi: 10.3389/fpsyg.2018.02571
- Williford, K., Rudrauf, D. & Landini, L. (2012). The paradoxes of subjectivity and the projective structure of consciousness. In S. Mingues & G. Preyer (Eds.), *Consciousness and subjectivity* (pp. 321-353). Ontos Verlag.

SHAUN GALLAGHER PhD, D.Phil (hon), is the Lillian and Morrie Moss Professor of Excellence in Philosophy at the University of Memphis. He has a secondary position as Professorial Fellow at the University of Wollongong. His areas of research include phenomenology, philosophy of mind, and embodied cognition.

Address: Department of Philosophy, Clement Hall 337, University of Memphis, Memphis, TN 38152. E-mail: s.gallagher@memphis.edu – ORCID: 0000-0002-3147-9929



SELF-CONSCIOUSNESS, Ψ AND Φ

(*Autoconsciencia, ψ y ϕ*)

Carlota Serrahima*

University of Valencia

<https://orcid.org/0000-0001-9525-6508>

Keywords

Self-consciousness
Psychological ownership
Bodily ownership
Immunity to error through
misidentification

ABSTRACT: This paper contributes to, and underpins, the project of an integrative account of psychological and bodily self-consciousness. Following a trend in the literature that uses the notion of “ownership” to discuss self-consciousness, and adapting José Luis Bermúdez’s terminology, I refer to psychological self-consciousness as “ ψ -ownership” and to bodily self-consciousness as “ ϕ -ownership.” The paper has two main aims. First, it presents a methodological framework for the study of ψ - and ϕ -ownership based on the framework that Bermúdez has put forward for this purpose. My framework involves three desiderata that apply *mutatis mutandis* to both the ψ and the ϕ domains. By laying out this framework, I clarify the relationship between ownership and the epistemic phenomenon of immunity to error through misidentification (both in the ψ and the ϕ domains), and I highlight the appeal of the notion of a phenomenology of ψ - and ϕ -ownership. Hence, this methodological discussion helps disclose deep similarities between the very explananda of the inquiries into ψ - and ϕ -ownership. Second, the paper critically examines one aspect of Bermúdez’s approach to self-consciousness, namely his explanation of bodily immunity to error in view of his defense of the phenomenology of bodily ownership. In order to articulate this criticism, I leverage the conclusions of the first half of the paper about ownership and immunity. I suggest that, despite Bermúdez’s explicit advocacy for the integrative project, his strategy at this specific point prevents him from seeing this project to completion.

Palabras clave

Autoconsciencia
Pertenencia psicológica
Pertenencia corporal
Inmunidad al error por identificación errónea

RESUMEN: Este artículo respalda y contribuye a una explicación integradora de la autoconsciencia psicológica y corporal. Siguiendo una tendencia en la literatura que emplea la noción de «pertenencia» (*ownership*) para analizar la autoconsciencia —y adaptando la terminología de José Luis Bermúdez—, en este artículo me refiero a la autoconsciencia psicológica como « ψ -pertenencia» y a la autoconsciencia corporal como « ϕ -pertenencia». El artículo tiene dos objetivos principales. En primer lugar, presenta un marco metodológico para el estudio de la ψ - y la ϕ -pertenencia, basado en el marco ya propuesto con anterioridad por Bermúdez. Mi marco incluye tres desiderata que se aplican *mutatis mutandis* a ambos dominios. Al exponer este marco, aclaro la relación entre esa noción de pertenencia y el fenómeno epistémico de la inmunidad al error por identificación errónea (tanto en el ámbito ψ como en el ϕ), y subrayo el atractivo de la postura según la cual existe una *fenomenología* de la ψ - y la ϕ -pertenencia. Esta discusión metodológica, por tanto, ayuda a revelar similitudes profundas entre nuestros dos explananda (esto es, la autoconsciencia psicológica y la corporal). En segundo lugar, el artículo examina críticamente un aspecto del enfoque de Bermúdez sobre la autoconsciencia: su explicación de la inmunidad al error en el caso corporal, en vista de su defensa de la fenomenología de la ϕ -pertenencia. Para articular esta crítica, retomo las conclusiones de la primera parte del artículo sobre la relación entre pertenencia e inmunidad. Sostengo que, a pesar del apoyo explícito de Bermúdez al proyecto integrador, la estrategia que adopta en este punto específico le impide llevar dicho proyecto a su plena realización.

* **Correspondence to:** Carlota Serrahima. Departament de Filosofia, Facultat de Filosofia i Ciències de l’Educació, Universitat de València. Avinguda Blasco Ibáñez, 30, El Pla del Real, 46010, València, Spain – carlota.serrahima@uv.es – <https://orcid.org/0000-0001-9525-6508>

How to cite: Serrahima, Carlota (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.27206>).

Received: 01 April, 2024; Final version: 17 September, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

Gako-hitzak

Autokontzientzia
Jabetza psikologikoa
Gorputz-jabetza
Identifikazio okerretik
sortutako errorearekiko
immunitatea

LABURPENA: Lan honek autokontzientzia psikologikoa eta gorputzaren autokontzientzia modu integratzailean azaltzeko proiektua aberastu eta sostengatuko du. Autokontzientzia aztertzeko «jabetza» nozioaz baliatzen den literaturako ildo bati jarraikiz, eta José Luis Bermúdezen terminologia egokituz, « ψ jabetza» deituko diot autokontzientzia psikologikoari, eta « ϕ jabetza», berriz, gorputzaren autokontzientziari. Bi helburu nagusi ditu lanak. Lehenik, marko metodologikoa bat aurkeztuko du ψ jabetza eta ϕ jabetza aztertzeko, Bermúdezek xede horretarako proposatu duen markoan oinarrituta. Hiru baldintza barne hartzen ditu nire markoak, eta *mutatis mutandis* aplikatzen zaizkie ψ eta ϕ eremuari. Esparru hori azaltzean, argi utziko dut zer harreman dagoen, batetik, jabetzaren eta, bestetik, identifikazio oker batetik eratorritako errorearen aurkako immunitatearen fenomeno epistemikoaren artean (bai ψ eremuan, bai ϕ eremuan), eta nabarmenduko dut interresgarria dela ψ eta ϕ jabetzen fenomenologiaren nozioa. Horrenbestez, eztabaida metodologikoa hori lagungarria da gure bi *explananda* (autokontzientzia psikologikoa eta gorputzaren autokontzientzia) horien arteko antzekotasun sakonak agerian jartzeko. Bigarrenik, lanak kritikoki aztertuko du Bermúdezek autokontzientziari buruz duen ikuspegiaren alderdi bat; hain zuzen ere, gorputzak errorearekiko duen immunitateari buruzko haren azalpena, ϕ jabetzaren fenomenologiaren haren aldekotasunaren argitan. Kritika hori mamitzeko, lanaren lehenengo zatian jabetzari eta immunitateari buruz ateratako ondorioetara itzuliko naiz. Nire iritzian, ikusiko dugunez, Bermúdezek proiektu integratzailea modu esplizituan defendatzen badu ere, puntu zehatz horretan darabilen estrategiak ez dio uzten proiektu hori bururaino eramaten.

1. Introduction: integrating ψ - and ϕ -ownership

In recent years, debates about self-consciousness in philosophy of mind have introduced the concept of ownership to help articulate otherwise traditional debates in this domain. Ownership talk has been found useful because, when subjects express self-consciousness, they often *self-attribute* certain entities widely taken to be essential to their nature as subjects: among such things, importantly, psychological states and properties on the one hand, and bodily states and properties on the other. Indeed, the shape that the debate on self-consciousness has traditionally taken is one in which we inquire into how subjects see themselves as bearers of psychological states and properties, and they also see themselves as embodied. José Luis Bermúdez (2018a) refers to these two sorts of ownership as *ψ -ownership* and *ϕ -ownership*, respectively. I assume his terminology in this essay.

Bermúdez is one of the voices in contemporary analytic philosophy of mind that has most clearly advocated for an integrative account of ψ - and ϕ -ownership. He has explicitly distanced himself from the tendency to make “[d]ebates about bodily ownership and psychological ownership [proceed] independently of each other” (*ibid.*, p. 233), and he has strived to develop explanations in each domain that mesh well with explanations in the other. The overarching aim of this paper is to contribute to, and to underpin, the project of an integrative account of ψ - and ϕ -ownership. On the one hand, the paper presents a methodological framework for the study of both ψ - and ϕ -ownership that refines the framework that Bermúdez himself (*ibid.*) has put forward for this purpose. On the other hand, the paper discusses one aspect of Bermúdez’s approach to self-consciousness that, I suggest, actually prevents him from seeing the integrative project to completion.

Current discussions on ψ - and ϕ -ownership have often revolved around the question of whether there is a *sense* of ownership: respectively, a *sense* that the experiences that we have are our own; or a *sense* that the body that we feel somatosensorily is our own.¹ But what *sense* means is rather unclear in this context. As it has been used, it sometimes points to some sort of *phenomenology*: respectively, a phenomenal mark that the mental states that we undergo are our own; or a phenomenal mark that the body that we feel is our own. Bermúdez, however, has rightfully pointed out that present-

¹ Debates on bodily ownership in philosophy have essentially focused on somatosensation as grounds for judgments. Somatosensation includes pain, proprioception, kinaesthesia, interoception, sensations related to balance, tactile sensations, and feelings of bodily temperature. In this paper I too focus on bodily ownership in somatosensation. Throughout the paper, I use “bodily sensation,” “bodily experience” and “somatosensation” interchangeably.

ing the discussion on ownership as if it is *the phenomenology* of ownership that needs to be accounted for “seems unwise” (*ibid.*, p. 237) —if only because it requires that participants in the debate take on, from the very outset, substantial commitments as to the existence of such phenomenological traits.

Another strategy for the study of ψ - and ϕ -ownership seems wiser. Even those who would rather not commit to the existence of a phenomenology of ownership will presumably agree that there is an interesting debate to be held around the fact that subjects often make *judgments* of ownership, both ψ - and ϕ -. Indeed, when subjects have conscious mental states or experiences, they typically report them by using the first-person pronoun “I” in the subject position, and attributing the mental state or experience to the subject. Examples of this are “I am sad” or “I am thinking about the storm” when declared on the grounds of the relevant experiences, namely, respectively, my feeling sad and my occurrent thought about the storm. In turn, when subjects feel sensations in a body, they typically report the sensations by using the first person pronoun “my” to refer to the body that they are feeling, thereby self-attributing it. This happens, for instance, in “my chin itches” or “my finger hurts” when declared, respectively, on the grounds of feeling the relevant sensations, namely an itch in the chin or a pain in the finger.²

A basic tenet of Bermúdez’s framework is that, when we inquire into ψ - and ϕ -ownership, it is *judgments* that need to be explained (respectively, judgments of ψ -ownership and judgments of ϕ -ownership, such as “I am sad” or “my chin itches”): judgments of ψ - and ϕ -ownership are our basic data. In this paper, I assume this tenet.

According to Bermúdez, the inquiry into ownership then consists in answering two questions, which apply to both types of judgments. On the one hand, a “descriptive-causal question about the source or origins of those judgments” (*ibid.*) —that is, a question about the mechanisms that underlie the production of the self-attributions, given the relevant grounds. On the other hand, “a normative question about the reasons for which those judgments are made” (*ibid.*) —a question about how judgments of ψ - and ϕ -ownership are justified by their respective grounds. Views on psychological ownership are then views that address these questions for judgments in which subjects self-attribute psychological states and properties on experiential grounds; and views on bodily ownership are views that address these questions for judgments in which subjects self-attribute bodily states and properties on somatosensory grounds. In turn, an *integrative* approach to ψ - and ϕ -ownership strives to address these questions in each of the domains, while making sure that the answers in one domain mesh well with the answers in the other.

In **Section 2** of this paper, I take up the task of outlining the research agenda into ψ - and ϕ -ownership in a way that assumes, but also refines, Bermúdez’s normative and causal-descriptive questions. My agenda is based on the three desiderata for views on ϕ -ownership that I laid out in Serrahima (2023), and which I succinctly recount in 2.1. Those desiderata, I submit, can be extended *mutatis mutandis* to views on ψ -ownership (as I detail in 2.2). On the basis of these desiderata, I make various things explicit. First, how the landscapes of existing views on, respectively, ψ - and ϕ -ownership, map onto each other. Second, how answers to Bermúdez’s normative and causal-descriptive questions need to be constrained, both in the ψ and ϕ domains, by parallel philosophical intuitions on the one hand, and by structurally similar empirical cases on the other. And third, how discussions on ψ - and ϕ -ownership relate to discussions about the epistemic properties of judgments of ownership in both the ψ and ϕ domains; in particular, to immunity to error through misidentification.

Hence, although Section 2 puts forward an essentially methodological point, my point there is not *merely* methodological: the methodological discussion —i.e. the definition of desiderata for views on ψ - and ϕ -ownership— helps disclose

² Alternative, natural expressions of what these two statements express would be “I can feel than my chin itches” and “I feel pain in my finger.” In these statements, the first-person concept is used, not only to refer to the subject of a body (part), but also to the subject of the sensations by which it is felt (through the italicised pronoun “I”). In other words, in judgments of ϕ -ownership, ψ -ownership is typically also present, even if sometimes only implicitly. I have argued in Serrahima (2024) that this is a very important feature of ϕ -ownership —one that speaks in favour of an integrative approach to ψ - and ϕ -ownership at that. Bermúdez (2018a, pp. 246-7) also appeals to the significance of this feature for the integrative approach. In this paper I want to provide other reasons in favour of the integrative approach, and hence I will leave this feature aside.

deep similarities between the very explananda of the inquiries into ψ - and ϕ -ownership. When inquiring into ψ - and ϕ -ownership, we follow similar explanatory routes, guided by similar intuitions and empirical cases, simply because we are dealing with two aspects of the same phenomenon, i.e. how subjects experience and think of themselves as such. In this sense, my discussion in Section 2 contributes to reinforcing the integrative approach to ψ - and ϕ -ownership.

In **Section 3** I move on to discussing one aspect of Bermúdez's treatment of self-consciousness specially developed in his "The Elusiveness Thesis, Immunity to Error through Misidentification, and Privileged Access" (2018b). As it turns out, Bermúdez's views on the nature of ψ - and ϕ -ownership are significantly different: while he defends that there is, indeed, a *phenomenology* of bodily ownership in bodily sensations (Bermúdez 2018c, 2018d, 2018e, 2020), he advocates for the so-called "Elusiveness Thesis" in the psychological domain. In brief this is the thesis that, leaving aside bodily sensations, in conscious mental states "the self cannot be an object of introspective awareness" (Bermúdez, 2018b, p. 97).

Section 3 examines this difference in Bermúdez's approach to ownership *vis à vis* his account of immunity to error through misidentification (IEM). IEM is an epistemic phenomenon that philosophers have identified both in the ψ and the ϕ domains: with some variations, psychological self-attributions on experiential grounds, and bodily self-attributions on somatosensory grounds, have both been found to be immune to error through misidentification. In his 2018b, Bermúdez offers an account of bodily IEM. He models this account on his view about psychological IEM, by proposing what he calls a *simple theory* in both the ψ and the ϕ domains (even if his view about psychological IEM, as we will see, is based on the Elusiveness Thesis). I will argue that Bermúdez could, much more convincingly, resort directly to his own commitment to the phenomenology of bodily ownership in his explanation of bodily IEM, especially given the explanatory relationship between IEM and the phenomenology of ownership as will be disclosed in Section 2. Not doing so, I will suggest, has the undesirable consequence that the phenomenology of bodily ownership appears as an epiphenomenon, explanatorily and causally inert.³ I will conclude the paper with a tentative diagnosis of the source of Bermúdez's strategy and a note on the prospects of the integrative project under the light of this diagnosis.

2. Senses of ownership

2.1. Explaining ϕ -ownership

I will start to lay out the agenda for an integrative approach to ownership by focusing, first, on ϕ -ownership. Consider judgments of ϕ -ownership, such as "my chin itches" or "my finger hurts," in which one singles out a somatosensorily felt body (part) of which one judges that it is one's own. Let us say that judgments of this sort express *awareness that the felt body is one's own*.⁴ An interesting question now is what this awareness consists in: what does it mean, exactly, to be aware of the body that one is feeling *as one's own*?

Two broad routes are explored on this count. The first says that all there is to this awareness is judgments of ϕ -ownership: bodily ownership is, so to speak, mainly a cognitive achievement, grounded on otherwise "self-less" bodily

³ I am indebted to Aarón Álvarez-González for this formulation of the problem.

⁴ This *awareness that the felt body is one's own* need not be factive. On the one hand, the body (part) in question may not exist, but subjects may *be aware of it as their own body (part)* in the relevant sense. This is often the case in phantom limb experiences. On the other hand, it is customary to assume that, in some science-fictional scenarios where one would be wired up to another body, one could be somatosensorily aware of a body (part) that does exist, but that is someone else's. Here too, as these science-fictional scenarios go, one would *be aware of the relevant body (part) as one's own* (Wright, 2012; de Vignemont, 2012; García-Carpintero, 2018. But Salje (2017) disputes the customary description of these science-fictional scenarios). I will briefly come back to this point in 2.3, in discussion of immunity to error through misidentification (see fn. 13).

experiences (Alsmith, 2015; Chadha, 2018). The second route, in contrast, says that there is a phenomenology of bodily ownership, namely that being aware of the body as one's own consists in *experiencing* the body as one's own. As we noted above with Bermúdez, the phenomenology of bodily ownership clearly plays the role of an explanans here, rather than being an assumption to make at the very outset of the discussion.

Most authors writing about ϕ -ownership take the phenomenological route—including Bermúdez. Their debate then revolves around how the phenomenology of bodily ownership should be specified. Together with Martin (1995), Bermúdez (2018c, 2018d, 2018e, 2020) has argued that the phenomenology of bodily ownership consists in the peculiar spatial content, and the corresponding spatial phenomenology, of bodily experiences. Others have appealed to affective phenomenology (de Vignemont, 2018; Bradley, 2021); agential phenomenology (Gallese & Sinigaglia, 2010; Peacocke, 2017; Gallagher, 2019); the peculiar presentation of bodily properties when they are somatosensorily felt (Dokic, 2003; Serrahima, 2024); the character of bodily sensations as Husserlian double-sensations (Lotan, 2022); or an irreducible mineness quale (Billon, 2017).⁵

In any case, Bermúdez's descriptive-causal and normative questions have warned us about a basic constraint one must bear in mind when opting for one route or the other. We ultimately need to explain why, in typical judgments about somatosensory states, one refers to the felt body first-personally. Hence, when inquiring into ϕ -ownership, we must see to characterise the contents and phenomenology of somatosensation, as well as the mechanisms by which one transitions from the sensations to the judgments, in a way that, on the one hand, makes it plausible that judgments with this specific structure are formed (as per the descriptive-causal question), and on the other hand, warrants the judgments (as per the normative question).

We can summarise this first part of our research agenda into ϕ -ownership in the following

Judgment Formation Goal (ϕ): accounts of ϕ -ownership need to explain judgments of bodily self-attribution made on somatosensory grounds. That is, they need to detail the mechanisms underlying the judgments' production, including the contents and phenomenology of the sensations; and they need to explain how the grounds justify the judgments.

Our research agenda, however, does not end here. The inquiry into ϕ -ownership, as dictated by the Judgment Formation Goal, needs some tailoring to the specific character of the transitions that we are dealing with, namely those from somatosensations to judgments grounded on them. There is one intuition about these transitions that has been crucial in debates on ϕ -ownership: the intuition that there is a *uniquely compelling link* between actually having bodily sensations, and the fact that, when we judge on them, we describe the felt body as *mine*. On some occasions, this intuition has even brought authors to claim that the link in question is (perhaps conceptually) necessary: indeed, some have claimed that it is just impossible that one feels sensations in a body, and yet one does not feel this body as one's own (Martin, 1995; Dokic, 2003; Bradley, 2021).

The role that this intuition of compellingness should play in the debate on ϕ -ownership is, I believe, an urgent matter of discussion in its own right.⁶ While the compellingness of the link has been a decisive background assumption in discussions on bodily self-consciousness, it is the case that some subjects have experiences in which the link between the instantiation of bodily sensation, and reports thereof that self-attribute the body, is broken, or at the very least tenuous.

⁵ Note that the question that gives structure to the debate (what does it mean, exactly, to be aware of the body that one is feeling *as one's own*?) can in principle be replied by characterising the *contents* of bodily sensations in first-personal terms, but also by seeing the phenomenology of ownership as contributed by the *mode of awareness*. Besides, to the extent that the non-phenomenological route is feasible, the question could also be replied by relying on the mode of awareness, but without committing to the mode having a phenomenological impact on the experiences.

⁶ The considerations that I make in the rest of this subsection about the intuition of compellingness nuance those in Serrahima (2023), which, after reading Jeppsson (2025), seem to me to overstate the importance of the intuition.

De Vignemont (2013) brought the case of somatoparaphrenic patients to bear specifically on this issue. Somatoparaphrenic patients have delusional beliefs about the contralesional side of their bodies according to which it does not belong to them (Invernizzi *et al.*, 2013). Yet, some of them are able to feel sensations in their “disowned” limbs. Furthermore, people who suffer from depersonalisation describe their experiences, for instance, as follows: “sometimes, I experience myself as much smaller; a little homunculus residing in the head of the full-sized body. I can still move it around with great mastery and ease, but I am now more comparable to an expert car driver or wheelchair user” (Jeppsson, 2025, p. 86). These kinds of bodily awareness do not neatly fit into the mould of full-blown ownership, and they call into question the intuition of compellingness.⁷

This dialectic shows that, when accounting for ϕ -ownership, we need to reach an equilibrium. On the one hand, we need to respond to what arguably lies behind the intuition of compellingness: that somatosensory experiences *typically* involve ϕ -ownership. In other words, that presumably for most subjects, most of the time, the body that they experience somatosensorily is presented to them as their own body. It might even be that, typically, the subjects who are prone to form judgments of ϕ -ownership when they have sensations, are prone to do so compellingly: this might explain why the alternative seems unconceivable to some, as often expressed in philosophical theorising. Note that, in this context, the notion of a phenomenology of bodily ownership has some plausibility. In general, one straightforward way to explain the formation of, and to justify, a subject’s belief that p , is by appeal to her experience that p . If a subject *experiences* the body as her own when she perceives it from the inside, then this explains and justifies the subject’s self-attributive judgments, and it constitutes a good explanation of their compellingness, since judging just involves taking experiences at face value.

However, on the other hand, views on ϕ -ownership also need to be empirically adequate. To the extent that the link between having a bodily sensation and self-attributing the felt body is not necessary, we must characterise bodily ownership in a way that makes it possible that bodily self-attributions be missing and the sensations still be there. This requirement is in place even if one subscribes to the thesis of a phenomenology of bodily ownership: it must be possible that the relevant phenomenal component be missing —i.e. the phenomenal component standing for bodily ownership— in some subjects’ bodily sensations.

The upshot is that our research agenda into ϕ -ownership needs to be guided by two further goals, which we can express succinctly thus:

Typicality Goal (ϕ): accounts of ϕ -ownership need to explain the fact that for most subjects, most of the time, bodily sensations yield judgments of bodily self-attribution.

Empirical Goal (ϕ): accounts of ϕ -ownership need to be compatible with reports of bodily sensation in which the subject does not self-attribute the felt body.

These two goals are auxiliary to the Judgment Formation Goal. With Bermúdez, we started claiming that judgments of ϕ -ownership are the basic data to be explained. In these judgments we express awareness that the felt body is our own; and we need to explain the nature of this awareness in a way that covers the causal-descriptive and normative aspects of judgment formation. The Typicality and Empirical Goals say that all of this needs to be extensionally adequate.

This concludes my agenda for the study of ϕ -ownership. This agenda covers Bermúdez’s normative and causal-descriptive questions, but it also fixes some constraints on their possible answers. I will now move on to showing that the agenda guiding the study of ϕ -ownership is also suitable to guide (and actually guides, *mutatis mutandis*) research on ψ -ownership.

⁷ Bradley (2021) objects to de Vignemont’s (2013) diagnosis about somatoparaphrenia and ϕ -ownership. He thematises specifically the intuition of compellingness, and he interprets somatoparaphrenia, as well as depersonalisation, as compatible with it. Jeppsson (2025), in turn, is critical with how philosophers interpret reports of psychopathological patients for their theoretical purposes, taking Bradley (2021) as a case in point.

2.2. Explaining ψ -ownership

When we discuss ψ -ownership, the relevant data to be explained are judgments in which subjects self-attribute phenomenally conscious experiences. To recall, examples of this would be, for instance, “I am sad” or “I am thinking about the storm” when judged on the grounds of, respectively, my feeling sad and my occurrent thought about the storm. Let us say that, when subjects make this kind of judgment, they express *awareness that the experiences they have are their own*.⁸ We can now wonder, as we did with ϕ -ownership, what this awareness consists in: what exactly is implied in *being aware of conscious experiences as one’s own*. Our question now does not concern ownership over the body as is presented in a subclass of conscious experiences, as was the case when discussing ϕ -ownership. Rather, it is more general: we are concerned with ownership over conscious experiences themselves.

Broadly speaking, two routes can be followed on this count, which draw a taxonomy of views very similar to that of views on ϕ -ownership. The first is the route of those who believe that the first person that figures in judgments of ψ -ownership does not have a phenomenal counterpart at the level of conscious experiences themselves. Some authors within this trend present arguments to the effect that our explananda, i.e. judgments of ψ -ownership, simply do not require that we posit such phenomenal counterpart (Praetorius, 2009; Schear, 2009; McClelland, 2023; Salje & Geddes, 2023; Howell, 2023; Soldati, 2023; Wu, 2023; Kang, 2025). For others, Humean intuitions that the self is not introspectable, or relatedly, metaphysical claims about the nature of the non-introspectable self, also speak against the phenomenology of ψ -ownership (Dainton, 2008; Prinz, 2011; Siderits, 2011; Chadha, 2018). In particular, Bermúdez himself takes in the Humean intuition that the self cannot be introspected, hence subscribing to the so-called *Elusiveness Thesis* (meaning that the self is introspectively elusive). Bermúdez’s version of the Elusiveness Thesis is based on Shoemaker’s (1986). On Shoemaker’s construal, in introspection we do not gather “information that allows [us] to pick out one object among a range of objects in virtue of its perceived properties” (Bermúdez, 2018b, p. 101) —more specifically, introspection does not involve

⁸ The question whether the awareness expressed in self-attributions of experiences can be non-factive is more complicated than the question about the non-factivity of awareness in bodily self-attributions. In the bodily case, the non-factivity of the awareness could come, as it were, from two places (see fn. 4). On the one hand, from the inexistence of the self-attributed object, i.e. the body (part). Phantom limbs are an example of this. A parallel case in the psychological domain might be possible: it might be true that one can seem to have an experience *E* that one does not actually have, where these seeming experience *E* is taken to be one’s own (say: it seems to me that I have a pain in the collar bone, *E*, of which I think and talk about as *my* collar bone pain. But, as it turns out, I don’t have a pain in the collar bone, but what I actually have is a toothache). We may want to describe this situation as one in which one has non-factive awareness *of E* as one’s own: *E* doesn’t actually exist, but one is aware of it as existing and being one’s own. On the other hand, in the bodily case, the non-factivity of the awareness may also come from the ownership side: it is customary to assume that, in some science-fictional scenarios, subjects could be somatosensorily aware of someone else’s (existing) body as one’s own (Evans, 1982, p. 221); in which case one’s awareness of the body as one’s own would be non-factive because, even if the body in question exists, it would not be one’s own. It is unclear that we can have a case that parallels this in the psychological domain. This would be a case in which I am aware of a phenomenally conscious experience *E* as *my own*, and the non-factivity consists in that, although the experience in question does exist, *it is in fact not my own*. The reason why it is unclear that this can happen is that, when it comes to experiences, phenomenal access to them arguably implies that one is in fact their owner, at least in a minimal sense (Coliva, 2002). This is so even if, in some cases, one has phenomenal access to thoughts that one is not straightforwardly inclined to self-attribute. Reports of patients with thought-insertion question the idea that, given a phenomenally conscious thought, one necessarily has a *sense* that the thought is theirs. This has indeed brought theorists to refine the analysis of self-consciousness in conscious thought, paradigmatically by introducing a distinction between the sense of ownership and the sense of agency over thoughts, which then allows to say that the sense of agency, but not the sense of ownership, is impaired in thought insertion (Stephens & Graham, 2000; Campbell, 1999; Campbell, 2002). To be clear, this concerns the phenomenological and epistemic side of things. On the factual side, however, there is less discussion: thought-insertion does not call into question (or at least not so clearly) that the phenomenally accessed thoughts *are in fact the subject’s own thoughts*, at least in the sense of being in the subject’s “own head,” even when it does not seem to the subject that they are (the experiences related in Jeppsson (2025), sections 3.2–3.5, for instance, appear to be in line with this). I will say a bit more about thought-insertion immediately in this section, when introducing the Empirical Goal (ψ).

picking out information whereby we individuate *ourselves* from other objects. I will come back to examining Bermúdez's commitment to the Elusiveness Thesis in the broader context of some of his other commitments about self-consciousness in Section 3 below.

Opposite to this approach to the nature of ψ -ownership are those who do advocate for a phenomenology of ψ -ownership. On their view, the awareness that conscious experiences are our own partly consists in *experiencing* that they are. Authors within this trend have worked on characterising the phenomenology of ψ -ownership, e.g. in terms of a first-personal "mode of givenness" of all conscious experiences (Zahavi, 2005; Fasching, 2009; Zahavi & Kriegel, 2015; Gallagher, 2015), in agentive terms (Duncan, 2019), in affective terms (Dub, 2023), or as a phenomenal mark jointly constituted by various phenomenal dimensions (Lane, 2012). Besides, this sort of position has often been defended by means of inferences to the best explanation, arguing that the best way to explain judgments of ψ -ownership and their features is by appealing to a phenomenology of ψ -ownership involved in their grounds (Grünbaum, 2012; Guillot, 2023; Nida-Rümelin, 2023).

Indeed, the choice of one view or another regarding the nature of ψ -ownership must be informed by the view's capacity to respond to Bermúdez's questions, now applied to judgments of ψ -ownership. We need to explain why, in typical judgments about experiences, one self-ascribes the experiences, using "I" in the subject position: what it is about conscious experiences themselves, or about the mechanisms by which one transitions from the experiences to the judgments, that makes it the case that judgments with this specific structure are formed (as per the descriptive-causal question); and warrants the judgments (as per the normative question). In this sense, inferences to the best explanation in defence of the phenomenology of ψ -ownership rely on the fact that, if there is such phenomenology, then the phenomenology motivates and justifies the judgments, which subjects just form by taking their experiences at face value.

It should be clear at this point that, in a nutshell, the agenda into ψ -ownership is guided by a desideratum that is essentially a twin of our first desideratum about bodily ownership:

Judgment Formation Goal (ψ): accounts of ψ -ownership need to explain judgments of self-attribution of experiences made on phenomenal grounds. That is, they must describe the mechanisms underlying the judgments' production, including the contents and phenomenology of conscious experiences; and they need to explain how the grounds justify the judgments.

The agenda into ψ -ownership does not end here, and neither does the parallel with bodily ownership. When we approach psychological ownership, our inquiry is, in some fundamental way, different from just any inquiry into how and why subjects claim some object or other to be their own; and it is different from that, in the same sense in which the inquiry into bodily ownership is different from that. Here too, the discussion is very much shaped by an intuition of compellingness: the intuition that there is a *uniquely compelling link* between phenomenal states, and judgments in which we self-attribute such states.

Again, the role that this intuition of compellingness must play in our theorizing about ψ -ownership is an urgent matter of discussion in its own right. That the intuition itself has been present in, and indeed has shaped, discussions on ψ -ownership is hardly disputable. One prime historical example of the influence of this intuition is Descartes' first item of knowledge, specially in his *Meditations on First Philosophy* (1641/ 2008). After arriving at a certainty about the occurrence of one type of mental activity, namely thought, Descartes famously claims that the proposition "I am, I exist" is true (Med. 2, AT 7:25). From this first item of knowledge then follows a reasoning about the nature of the I "that now necessarily exists" (*ibid.*). Within the Cartesian framework, the claim "I am, I exist" rests on arguments about the need for a substance that bears the attributes we become aware of—in this case, thoughts (Bermúdez, 2018e, pp. 99–100). But this is not an explicit part of the reasoning at this point of the *Meditations*. In the *Meditations*, the dialectic strategy consists in walking the reader from intuitive opinions to indubitable principles, the first of which is "I am, I exist." The connection between being phenomenally aware of mental states, and attributing the mental states to *me*,

which is the precursor to his existence claim,⁹ is presented by Descartes as a natural step arguably relying on an intuition that he assumes will be shared by his readers.¹⁰

Despite the force of this intuition, however, we now know that views on ψ -ownership need to reach an equilibrium, just as happened with views on ϕ -ownership. On the one hand, they need to answer to what arguably underlies the intuition of compellingness: that undergoing phenomenally conscious experiences *typically* involves ψ -ownership. Most subjects, most of the time, feel compelled to self-attribute the experiences they live through; or, in other words, are aware that the experiences they live through are their own. On the other hand, however, views on ψ -ownership also need to be empirically adequate. As a matter of fact, the link between the occurrence of conscious experiences and the self-attribution of these experiences by their subject is not in place in all cases. As Campbell (1999; 2002) points out, thought insertion bears directly on the discussion on ψ -ownership. In thought insertion, subjects have experiences that are not straightforwardly presented to themselves as their own: they might claim that certain thoughts they are conscious of are not theirs, or that they are someone else's (see Frith, 1992). Hence, in the domain of ψ -ownership too, however we choose to explain the first person figuring in judgments of ownership, we need to appeal to something that still makes it possible that, at least in some cases, self-attributions be missing and the experiences still be there, even if maybe altered.

Two further desiderata for our research agenda into ψ -ownership capture this dialectic. They qualify the first desideratum, calling attention to the fact that our explanations of psychological self-attributions need to be extensionally adequate. In turn, these desiderata are parallel to the second and third desiderata for ϕ -ownership:

Typicality Goal (ψ): accounts of ψ -ownership need to explain the fact that for most subjects, most of the time, phenomenally conscious experiences yield self-attributions of these experiences.

Empirical Goal (ψ): accounts of ψ -ownership need to be compatible with reports of phenomenally conscious experiences in which the subject does not self-attribute the experiences.

In sum, throughout 2.1 and 2.2 I have outlined a research agenda into ψ - and ϕ -ownership that not only spells out Bermúdez's normative and causal-descriptive questions, but also shows how research into both ψ - and ϕ -ownership has been guided by parallel intuitions, and must be sensitive to structurally similar cases that countervail the significance of these intuitions.

In the remainder of this section (2.3), I will focus on immunity to error through misidentification (IEM). I am particularly interested in a relatively high-level aspect of the discussion on IEM, and I will not engage in depth with the intricacies of any particular, ground-level view on the phenomenon. I am interested in the fact that the discussion on IEM is articulated, at least partly, around the intuitions of compellingness about ψ - and ϕ -ownership mentioned in the foregoing sections. This fact helps illuminate how ψ - and ϕ -ownership relate to IEM; and more specifically, it helps see why the notion of a phenomenology of ownership, both in the ψ and ϕ domains, is a *prima facie* good candidate to explain, respectively, psychological and bodily IEM. Relative to the overall objectives of the paper, the next subsection serves two purposes. First, it is the last piece of my not-merely-methodological point: it rounds up my case for the integrative approach to ψ - and ϕ -ownership, by pointing out that the parallelisms between the two domains extend to the epistemology of judgments of ownership. Second, it is an important building block for my critical comment on Bermúdez in the last part of the paper.

⁹ The text of the *Meditations* reads thus: “[C]ertainly I did exist, if I convinced myself of something. —But there is some deceiver or other, supremely powerful and cunning, who is deliberately deceiving me all the time.— Beyond doubt then, I also exist, if he is deceiving me; and he can deceive me all he likes, but he will never bring it about that I should be nothing as long as *I think* I am something” (Med. 2, AT 7:25. The translation is from 1641/ 2008. The emphasis is mine). The “*I think*” italicised in this passage is the self-attributive precursor I am alluding to in the body of the text. This precursor is actually more explicit in Descartes’ *Discourse on the Method*, in which the formulation of the first item of knowledge is his famous “*I think, therefore I am.*”

¹⁰ Classical discussions of this move are Lichtenberg’s notes in his *Waste Books* (2012, p. 152) and Russell (1945).

2.3. Explaining immunity

Shoemaker (1968) introduced the notion of immunity to error through misidentification in order to capture a specific kind of epistemic security of some of our singular judgments of the form “a is F.”¹¹ A judgment “a is F” is immune to error through misidentification (IEM) when it cannot be the case that the judgment is mistaken in the specific sense that the subject knows of *some* individual that they are F, but is wrong in predicating F *of the particular individual* picked out with “a.” Judgments that have this property are IEM relative to the use of “a.”

In this domain, one of the most discussed types of judgments is judgments in which one predicates something of oneself, namely judgments of the form “I am F.” Following the scheme just mentioned, a first-personal judgment “I am F” is IEM relative to “I” if it cannot be the case that the judgment is wrong because, although one *knows* that *someone* is F, one is mistaken in judging that the individual that one knows to be F is *oneself*.

What is most interesting about IEM is that it is a feature of judgments relative to their grounds. Consider the judgment “I am sad,” in which one predicates of oneself the psychological property of being sad; and suppose that one judges so on the grounds of having, phenomenologically, a feeling of sadness. In other words, consider the situation in which “I am sad” is a judgment of ψ -ownership —one made on the grounds of experiencing sadness, conveying awareness that the sadness is one’s own. This judgment is IEM, on Shoemaker’s view: phenomenal awareness¹² of psychological properties is such that, if one judges on its grounds that one is F, then the judgment won’t be wrong in the specific sense that, indeed, someone is F, and one *knows* that someone is F, but one is wrong in believing that it is oneself who is F.

This implies a substantive claim about phenomenal awareness as a way of accessing psychological properties; in particular, a claim about phenomenal awareness relative to the epistemic status of the judgments it yields. For contrast this with a case in which one judges “I am sad” on different grounds: suppose, for instance, that I judge “I am sad,” not because I feel sadness, but because I see an individual that looks very much like me, and who indeed looks sad, in what I think is a mirror. I then conclude “I am sad” on the grounds of reflecting about what I think are my looks. Unlike phenomenal awareness, these grounds —i.e. reflection on someone’s looks— leave open the possibility that the judgment is mistaken through misidentification. For suppose that, unbeknownst to me, it is not a mirror I am looking at. Rather, I just saw my identical twin sister, who is indeed very sad, through a glass wall; and then I, who was otherwise happy, have mistaken her for my own reflection and have concluded that I am actually sad. In this situation, my judgment is wrong exactly in the sense that, although someone is indeed sad, it is not me who is sad.

After Shoemaker, Evans (1982) famously pointed out that IEM relative to the first person is not limited to self-attributions of psychological properties on phenomenal grounds. Rather, Evans argued, bodily self-attributions are also IEM *if made on somatosensory grounds*. To put it in our terms, Evans’ point is that judgments of ϕ -ownership, in which subjects use the first person to refer to the body they are feeling from the inside, are also IEM.¹³ For an example, take the

¹¹ Strictly speaking, Shoemaker talked about statements. Following the convention in the literature, I will talk about judgments, on the assumption that statements express judgments.

¹² Or “introspection,” as participants in this discussion most often call it. I will use both terms interchangeably. I will also talk interchangeably about awareness and self-ascription of psychological properties, of psychological states, and of experiences.

¹³ As noted in fn. 4 (and also fn. 8), it is customary in the literature about bodily IEM to discuss cases in which subjects are wired up to someone else’s body that they feel is their own. On the grounds of these possible cases, it has been argued that bodily self-attributions on somatosensory grounds are only *de facto* IEM (Wright (2012, p. 272), de Vignemont (2012, p. 226), Guillot (2014, fn.7, 8 & 19), García-Carpintero (2018, pp. 19–20)): in these scenarios, the idea goes, the subject would know that some bodily property is instantiated, but she would be wrong that *she herself* instantiates it. But these scenarios are not usual in our world. In Serrahima (2025) I have discussed the *de facto* proviso on bodily IEM in comparison with psychological IEM, which is said to be logical. I have argued that this contrast has to do with differences in the metaphysics of bodily ownership *vis à vis* the metaphysics of psychological ownership. That discussion goes beyond the scope of the current paper. Here, the fact that intuitions about bodily IEM hold despite the possibility that bodily IEM be only *de facto* is grist for my mill. As I will suggest in this section, part of what lies behind IEM are the intuitions of compellingness about ownership. These intuitions are in place both in the psychological and bodily domains even if, in the bodily case, there may be worlds in which

judgment we would express as “My chin itches,” and suppose the subject judges so on the grounds of feeling an itch in the chin. In this situation, it is not possible that the subject knows that someone’s chin itches, but yet she is wrong in that the individual whose chin itches is in fact not herself. As Shoemaker’s view, Evans’ implies a substantive claim about grounds —somatosensory grounds in this case—, relative to the epistemic status of the judgments they yield: somatosensations are such that, if one accesses the body and its properties through them, and one judges e.g. “My chin itches,” then this judgment will not be wrong in the specific sense that someone’s chin itches but it is not mine.

The IEM feature of judgments of ψ - and ϕ -ownership thus point to a similarity between phenomenal awareness as grounds for judgments for ψ -ownership, and somatosensations as grounds for judgments of ϕ -ownership. One obvious similarity is that both capacities typically inform us about properties and states (psychological or bodily) that are *in fact* our own properties and states —which then makes it the case that, if we self-attribute these properties and states, we are typically not wrong. But this is not the similarity I am interested in here. The similarity I am interested in brings us back to the intuitions of compellingness highlighted in the previous subsections.

The IEM feature of judgments of ψ - and ϕ -ownership prescribes a condition on grounds, relative to the kind of knowledge they afford: if the subject knows on grounds G that someone has property F, then she *eo ipso* knows on the same grounds that she herself has property F. For the sake of clarity, let me spell out explicitly how this condition looks when applied, first, to phenomenal awareness and psychological properties:

If the subject knows on phenomenal grounds that someone has psychological property P, then she *eo ipso* knows on the same grounds that she herself has psychological property P.

And second, how the condition looks when applied to somatosensation and bodily properties:

If the subject knows on somatosensory grounds that someone has bodily property B, then she *eo ipso* knows on the same grounds that she herself has bodily property B.

What I want to highlight here is the importance of the “*eo ipso*” involved in (the two versions of) this condition. The interest of IEM as an epistemic phenomenon does not lie only on the factual side of things —i.e. on how my *knowledge* that the picked-up properties are mine depends on these properties actually being mine. The interest of IEM as an epistemic phenomenon also lies on how the link between picking out certain properties, psychological or bodily, on certain grounds, and being aware of these properties as my own, appears to be a very tight link. This is what the “*eo ipso*” above means: it highlights the unique compellingness of, on the one hand, taking psychological properties (or experiences) to be mine when I am aware of them phenomenally; and on the other hand, taking bodily properties to be mine when I pick them out somatosensorily. I contend that our epistemic intuitions about the IEM of judgments of ψ - and ϕ -ownership are driven by these intuitions of compellingness, and part of the point of the inquiry into the IEM of judgments of ψ - and ϕ -ownership is to figure out the extent to which this intuited compellingness makes a contribution to the privileged epistemic status of the judgments.

I believe that one possible way of seeing Shoemaker’s and Evans’ (otherwise diverse) approaches to IEM —respectively, psychological and bodily IEM— is as an implementation of this idea. On their view, part of the explanation of IEM relative to the first person lies on the fact that certain uses of “I” *do not rest on an identification*, in the sense of a recognition, of oneself as the individual of whom one predicates the relevant property. According to Shoemaker, in introspection there is no gap between having information about the instantiation of a given psychological property, and it appearing to one that one is the subject of this property (Shoemaker, 1968, pp. 563-5). In this framework, this is constitutively linked to the Elusiveness Thesis (for encountering the self in introspection could only have the function of

they do not track facts about ownership. The way I present IEM in the following paragraphs as “prescribing a condition on grounds” is essentially the same as in Serrahima (2025).

identifying the self as the owner of introspectable mental states, which is not required; Bermúdez, 2018b, p. 102). In turn, when it comes to bodily properties and somatosensory grounds, Evans writes that “for [one] to have, or appear to have, the information that the [bodily] property [F] is instantiated just is for it to appear to [one] that [one] is F” (1982, p. 221). That is to say, there is also no gap between being somatosensorily aware of the instantiation of a given bodily property, and being aware that one is the subject of this property. The no-identification route in these seminal approaches to IEM thus seems to articulate the intuitions of compellingness at the level of grounds —and to actually do so in the strongest way possible: it is not just that one compellingly self-attributes the properties picked out on certain grounds; it is rather that, whenever one picks out a given (psychological or bodily) property on the relevant grounds, one is *by the same token* aware of this property as one’s own property.

Unsurprisingly then, the empirical cases that question the tightness of the link between picking out properties and self-attributing them have been analysed in the literature as putative threats to the IEM thesis, both in the psy and phy domains.¹⁴ To recall, in thought insertion subjects have thoughts that are not presented to themselves straightforwardly as their own thoughts. In turn, patients with somatoparaphrenia deny ownership of bodily parts that they can feel somatosensorily. As explained in the previous subsections, these cases urge to nuance the intuitions that seem to have guided the debates: for most subjects, most of the time, given the grounds for the self-attributions, the self-attributions follow compellingly; but things are otherwise in thought insertion and somatoparaphrenia. These cases are more challenging to the IEM thesis, the more we find IEM to depend on the compellingness of the link between the awareness of properties on certain grounds, and the self-attribution of these properties.

Against this backdrop, we can now see two things about the relationship between the theoretical debates on ownership and immunity. The first is that explaining IEM partly relies on offering a view on ownership. More specifically, explaining psychological IEM partly relies on explaining the workings of ψ -ownership, and in particular the typical compellingness of judgments of ψ -ownership given conscious experiences; and in turn, explaining bodily IEM partly relies on explaining the workings of ϕ -ownership, and in particular the typical compellingness of judgments of ϕ -ownership given somatosensations. In this sense, the parallelisms between the ψ and ϕ domains extend from ownership to the epistemology of first-personal judgments. We can express this in a compressed way by slightly revising our Judgment Formation Goals, expressed jointly here:

Judgment Formation Goal (ψ & ϕ): accounts of ψ - and ϕ -ownership need to explain judgments of self-attribution (respectively, of experiences and of the body) made on the relevant grounds. That is, they must describe the mechanisms underlying the judgments’ production, including the contents and phenomenology of the relevant experiences; and they need to explain how the experiences justify the judgments, as well as other epistemic features of the judgments such as IEM.

Hence, the capacity of any given view of ownership, ψ - or ϕ -, to contribute to explaining IEM, is abductive reason in its favor.

The second thing to note about the relationship between ownership and immunity is that, to the extent that the existence of a *phenomenology* of ownership (ψ - or ϕ -) constitutes a good explanation of ownership, then the existence of a phenomenology of ownership (ψ - or ϕ -) is a *prima facie* good candidate to explain IEM (respectively, psychological or bodily). As said above, if there is a phenomenology of ownership, then subjects arguably make self-attributive judgments just by taking their experiences at face value, and this easily explains why they do so compellingly.

Of course, the notion of a phenomenology of ownership will be more plausible if we can adduce independent reasons in its favour, beyond explanatory capacity. Importantly, in this sense, things are not equally straightforward in the ψ and the ϕ domains. As the taxonomy of views sketched above reflects, while it is relatively common to defend that there is

¹⁴ See e.g. Campbell (2002); Coliva (2002); Hu (2017); Palmira (2020) for discussion of thought insertion and ψ -IEM; and Liang & Lane (2009); Rosenthal (2010); Kang (2016); de Vignemont (2012) for discussion of somatoparaphrenia and ϕ -IEM.

a phenomenology of *bodily* ownership, it is more controversial to take the phenomenological route in the psychological domain. It is not difficult to see why. Indeed, it is quite uncontroversial that in somatosensation subjects perceive a particular object, the body, as physically bounded, delimited with respect to other physical objects (Bermúdez 2018c, 2018d, 2018e). We may then argue that subjects perceive this particular object first-personally. But for introspection in general, the Elusiveness Thesis is popular, according to which we do not even get to individuating ourselves in introspection.

José Luis Bermúdez is an author who takes opposite routes on that count. While he has extensively advocated for the phenomenology of bodily ownership (*ibid.*), he defends the Elusiveness Thesis in the psychological domain. In the upcoming Section 3, I will focus on Bermúdez's approach to bodily IEM. Despite taking the opposite routes just mentioned, Bermúdez seems inclined to model his explanation of bodily IEM on his account of psychological IEM, even if the latter is based on the Elusiveness Thesis. More specifically, Bermúdez shies away from resorting to the phenomenology of bodily ownership, which is readily available in his own framework, to explain bodily IEM. The risk of this move is that the phenomenology of bodily ownership then appears as causally and explanatorily idle, given the relationship between ownership and immunity as described in this section. I will close Section 3, and the paper, with a tentative diagnosis of the sources of Bermúdez's move, followed by a note on certain theoretical choices we need to make if we want to fully honour the integrative project.

3. *A Bermúdez-inspired commentary on Bermúdez*

In his 2018b, Bermúdez offers a very insightful analysis of the relationship between the Elusiveness Thesis and IEM. Bermúdez highlights how the elusiveness of the self and the epistemic privilege of judgments of ψ -ownership may be “two sides of the same coin:” the epistemic privilege that we enjoy with respect to certain claims about our own psychological properties plausibly derives from the fact that our introspective self-awareness is awareness of a fundamentally different type than our awareness of other objects and people (*ibid.*, p. 97). In particular, in the Shoemakerian framework that Bermúdez adopts, IEM derives from the fact, explained above, that we do not pick out the self in introspection, and hence that introspection does not involve an identification of oneself. Things being thus, Bermúdez notes, one should then explain the fact that there are other forms of self-awareness that are also epistemically privileged, but yet not introspective. One of these is, of course, bodily sensation. Bodily sensation does involve perception of ourselves, but it also offers us privileged, indeed IEM, information about the body. Bermúdez concludes that, in this case, IEM requires an explanation “in which the elusiveness thesis has no part to play” (*ibid.*, p. 98). Here I want to focus precisely on how Bermúdez details this explanation. Although the Elusiveness Thesis has “no part to play” in it, Bermúdez borrows for this explanation some elements from his treatment of ψ -ownership and psychological IEM, in a way that, I contend, is not innocuous.

In discussion of ψ -ownership, once one subscribes to the Elusiveness Thesis, a fair question arises about how subjects actually make judgments of the form “I am ψ -ing” in the absence of any awareness, in experience, that they are *themselves* ψ -ing (Bermúdez, 2018b, p. 115). Bermúdez proposes that this transition results from the fact that subjects have “some level of mastery” of the link between being introspectively aware of a thought, and the fact that the thought is their own thought (*ibid.*).¹⁵ This is what he calls a *simple theory of introspection* (Bermúdez, 2018f). This mastery relies on the subjects' *appreciation* of the actual connection between the psychological properties they are aware of, and the fact that they are instantiated by themselves (*ibid.*, p. 92; 2018b, p. 115).

¹⁵ Bermúdez notes that the link between a subject being introspectively aware of a thought, and the fact that the thought is in fact the subject's own, is an *a priori* link (see also Peacocke, 1999), for reasons along the lines of those discussed in fn. 8.

Note that the question just mentioned, concerning the transitions from experiences to judgments of the form “I am ψ -ing,” sums up the causal-descriptive and normative questions involved in our Judgment Formation Goal (ψ). The simple theory of introspection aims to respond to these questions. A preliminary point that I want to make here is that I don’t find the simple theory of introspection satisfactory. Subjects are supposed to master the existing link between the instantiation of the psychological property and the fact that it is their own property, because they *appreciate* this existing link. But the notion of appreciation at stake here is evasive. It cannot mean that subjects are sensitive to the link in the sense that the link is somehow part of the phenomenology of their experiences—for this is in principle precluded by the Elusiveness Thesis. But still, the proposal that subjects appreciate the link should add something to the basic statement that, indeed, subjects respond to the link, finding the transition from their experiences to judgments of ψ -ownership compelling—for this is exactly what we aim at explaining. If the simple theory of introspection does not add something to this basic statement, it appears dangerously close to a redescription of our explanandum, and falls short of a causal explanation. It is unclear what the midway between these two undesirable options is.

Let us now turn to somatosensation. As mentioned, it is central to Bermúdez’s argumentation that the Elusiveness Thesis does not apply to somatosensation, for we do perceive ourselves by perceiving the body. Furthermore, in normal circumstances, the properties one feels from the inside are properties of one’s own body, and in this sense we can talk about an existing *link* between being aware of a bodily property in bodily sensation, and this property being a property of one’s own body.¹⁶ Because this link exists in normal circumstances, Bermúdez suggests, it is then plausible to defend a *simple theory of somatosensation*. To be clear, this is a theory about how subjects predicate of themselves the bodily properties they perceive: in particular, a theory partly modelled on the simple theory of introspection. According to the simple theory of somatosensation, then, subjects presumably have *some level of mastery* of the link between feeling a body and the fact that the felt body is theirs, because they *appreciate* this existing link.

It is however surprising that, at this particular point in his theorising, Bermúdez resorts to the simple theory, which in its original formulation for introspection relied on a notion of appreciation that, I have argued, is evasive. Bermúdez admits that “[o]ne would expect this simple theory [of somatosensation] to be more complicated than that implicated in ordinary psychological introspection, given the ... role of [somatosensation] in underwriting an awareness of one’s body as a bounded physical object uniquely responsive to the will” (2018b, p. 116). Indeed, in several of his writings Bermúdez has developed a detailed view about the peculiar spatial contents of bodily sensations: in bodily sensations, bodily properties are experienced within a bounded, body-shaped volume; and they are suitably connected to one another, as nodes of a network of further bodily locations.¹⁷ Besides, he has argued, these spatial contents have an impact on the phenomenology of somatosensation: there is a *phenomenology* of bodily ownership consisting precisely in spatial phenomenology (Bermúdez, 2018d).

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.

¹⁶ In the somatosensory case, the link between being aware of (bodily) properties, and the properties being properties of one’s own body, is only a *de facto* link, for the reasons discussed throughout fns. 4, 8, and 13.

¹⁷ For various formulations and elaborations of these ideas, see e.g. Bermúdez (1998), Bermúdez (2018c), Bermúdez (2018d), Bermúdez (2018e), and Bermúdez (2020).

It could be said that I am picking up on a merely verbal issue here: the proposal might be that, in introspection, subjects *appreciate* the link between (ψ) properties and themselves in one way; and in somatosensation, they *appreciate* the link between (ϕ) properties and themselves in another way, i.e. phenomenologically. I think the issue is significant, however. If there is a phenomenology of ϕ -ownership, but this phenomenology is not clearly embraced as an explanatory touchstone with respect to judgments of ϕ -ownership and bodily IEM, then it is explanatorily and causally idle; and resorting to the notion of a *simple theory* of somatosensation can indeed be read as shying away from clearly conceding a fundamental explanatory role to the phenomenology. Populating mental ontology with phenomenal ingredients that have explanatory potential but which do not actually explain what they could explain is to that extent unsatisfactory.

It is quite possible that Bermúdez can readily respond to the worry just expressed. However, I still think that it is fruitful to reflect on the possible reasons for the apparent disconnect, in his framework, between the phenomenology of bodily ownership and the explanation of bodily IEM. My speculation is this. Suppose one has found the Elusiveness Thesis convincing, and linked to the explanation of the IEM of judgments of ψ -ownership in terms of the absence of a recognition of the subject, as Bermúdez does. Then, one might feel a tension in accepting the exact opposite idea for ϕ -ownership: namely, that somatosensation involves a phenomenology of ϕ -ownership, and that this helps explain the IEM of judgments of ϕ -ownership in terms of the absence of a recognition of the subject. By accepting the latter idea, one would be simultaneously defending the elusiveness of the self, and first-person phenomenology, as accounting for the same phenomenon, *mutatis mutandis*, in different domains. These two explanatory routes are in principle not jointly incoherent, but they are in tension. This tension is specially pressing if we take into account the broader project we started with, namely that of explaining ψ - and ϕ -ownership integratively.

In the last sections of his 2018b, Bermúdez gives some further details about his understanding of the integrative project. He offers reasons to think that ψ - and ϕ -ownership are interdependent, and he suggests that the common thread running between them ultimately has to do with practical reasoning, since practical reasoning is reasoning about what one can do as an embodied subject, in which thought deliverances are constrained by the body's structure.¹⁸ However, the integrative project we aspired to is not satisfied solely by pointing to how ψ - and ϕ -ownership are closely interlinked, nor by identifying one paradigmatic area of interaction between them (such as practical reasoning). As uncovered throughout the paper, our integrative ambitions concerned causal-descriptive and normative aspects of the relationship between experiences and judgments of ψ - and ϕ -ownership, and the possibility of structural commonalities between ψ - and ϕ -ownership at the explanatory level. At this point we should wonder whether fulfilling these ambitions is really compatible with defending both that there is a phenomenology of bodily ownership, and that the self is elusive in introspection; and we should be ready to evaluate which one of these two commitments it would be less costly to drop.¹⁹

Acknowledgments

For discussion of ideas related to those reflected in this paper, I am grateful to the audience at the 30th SIUCC Conference in València. For fruitful discussion about earlier versions of the paper, and for their very useful comments, I am thankful to Aarón Álvarez-González, Francesco Consiglio, Manolo Martínez, Aida Roige, and two anonymous referees for this journal.

¹⁸ See also Bermúdez (2020) for a development of this idea.

¹⁹ I myself have defended that there is a phenomenology of bodily ownership and lean towards dropping the Elusiveness Thesis in introspection, but developing this is a matter for another paper.

REFERENCES

- Alsmith, A. (2015). Mental activity & the sense of ownership. *Review of Philosophy and Psychology*, 6(4), 881-896.
<https://doi.org/10.1007/s13164-014-0208-1>
- Bermúdez, J. L. (1998). *The paradox of self-consciousness*. MIT Press.
- Bermúdez, J. L. (2018a). Bodily ownership, psychological ownership, and psychopathology. In *The bodily self: Selected essays* (pp. 233-256). MIT Press.
- Bermúdez, J. L. (2018b). The elusiveness thesis, immunity to error through misidentification, and privileged access. In *The bodily self: Selected essays* (pp. 97-124). MIT Press.
- Bermúdez, J. L. (2018c). The phenomenology of bodily awareness. In *The bodily self: Selected essays* (pp. 125-154). MIT Press.
- Bermúdez, J. L. (2018d). Bodily ownership, bodily awareness, and knowledge without observation. In *The bodily self: Selected essays* (pp. 183-202). MIT Press.
- Bermúdez, J. L. (2018e). Ownership and the space of the body. In *The bodily self: Selected essays* (pp. 203-232). MIT Press.
- Bermúdez, J.L. (2018f). The sources of self-consciousness. In *The bodily self: Selected essays* (pp. 77-96). MIT Press.
- Bermúdez, J. L. (2020). Four theses about self-consciousness and bodily experience: Descartes, Kant, Locke, and Merleau-Ponty. *Journal of the American Philosophical Association*, 6(1), 96-116.
- Billon, A. (2017). Mineness first: Three challenges to the recent theories of the sense of bodily ownership. In F. de Vignemont & A. J. Alsmith (Eds.), *The subject's matter: Self-consciousness and the body* (pp. 189-216). MIT Press.
- Bradley, A. (2021). The feeling of bodily ownership. *Philosophy and Phenomenological Research*, 102(2), 359-379.
<https://doi.org/10.1111/phpr.12657>
- Campbell, J. (1999). Schizophrenia, the space of reasons, and thinking as a motor process. *The Monist*, 82(4), 609-625.
- Campbell, J. (2002). The ownership of thoughts. *Philosophy, Psychiatry, & Psychology*, 9(1), 35-39.
- Chadha, M. (2018). No-self and the phenomenology of ownership. *Australasian Journal of Philosophy*, 96(1), 14-27.
<https://doi.org/10.1080/00048402.2017.1307236>
- Coliva, A. (2002). Thought insertion and immunity to error through misidentification. *Philosophy, Psychiatry, & Psychology*, 9(1), 27-34.
- Dainton, B. (2008). *The phenomenal self*. Oxford University Press.
- Descartes, R. (1641/ 2008). *Meditations on first philosophy. With selections from objections and replies*. Oxford University Press.
- De Vignemont, F. (2012). Bodily immunity to error. In S. Prosser & F. Recanati (Eds.), *Immunity to error through misidentification: New essays* (pp. 1-27). Cambridge University Press.
- De Vignemont, F. (2013). The mark of bodily ownership. *Analysis*, 73(4), 643-651.
- De Vignemont, F. (2018). *Mind the body: An exploration of bodily self-awareness*. Oxford University Press.
- Dokic, J. (2003). The sense of ownership: An analogy between sensation and action. In J. Roessler & N. M. Eilan (Eds.), *Agency and Self-Awareness: Issues in Philosophy and Psychology* (pp. 320-344). Oxford University Press.
- Dub, R. (2023). Emotions of mineness. In M. García-Carpintero & M. Guillot (Eds.), *Self-Experience: Essays on Inner Awareness* (pp. 291-313). Oxford University Press.

- Duncan, M. (2019). The self shows up in experience. *Review of Philosophy and Psychology*, 10(2), 299-318. <https://doi.org/10.1007/s13164-017-0355-2>
- Evans, G. (1982). *The varieties of reference*. Oxford Clarendon Press.
- Fasching, W. (2009). The mineness of experience. *Continental Philosophy Review*, 42(2), 131-148. <https://doi.org/10.1007/s11007-009-9107-z>
- Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Lawrence Erlbaum Ass. Publishers.
- Gallagher, S. (2015). Relations between agency and ownership in the case of schizophrenic thought insertion and delusions of control. *Review of Philosophy and Psychology*, 6(4), 865-879. <https://doi.org/10.1007/s13164-014-0222-3>
- Gallagher, S. (2019). The senses of a bodily self. *Protosociology: An International Journal of Interdisciplinary Research*, 36, 411-431.
- Gallese, V., & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia*, 48(3), 746-755.
- García-Carpintero, M. (2018). De se thoughts and immunity to error through misidentification. *Synthese*, 195(8), 3311-3333. <https://doi.org/10.1007/s11229-015-0817-y>
- Grünbaum, T. (2012). First-person and minimal self-Consciousness. In S. Miguens & G. Preyer (Eds.), *Consciousness and subjectivity* (pp. 273-296). Ontos Verlag.
- Guillot, M. (2014). Identification-free at last: Semantic relativism, Evans's legacy and a unified approach to immunity to error through misidentification. *Teorema: International Journal of Philosophy*, 33(3), 7-30.
- Guillot, M. (2023). The phenomenal concept of self and first-person epistemology. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 223-249). Oxford University Press.
- Howell, R. J. (2023). Transparency and subjective character. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 77-98). Oxford University Press.
- Hu, I. (2017). The epistemology of immunity to error through misidentification. *The Journal of Philosophy*, 114(3), 113-133.
- Invernizzi, P., Gandola, M., Romano, D., Zapparoli, L., Bottini, G., & Paulesu, E. (2013). What is mine? Behavioral and anatomical dissociations between somatoparaphrenia and anosognosia for hemiplegia. *Behavioural Neurology*, 26(1-2), 548467. <https://doi.org/10.3233/BEN-2012-110226>
- Jeppsson, S. (2025). Allegedly impossible experiences. *Philosophical Psychology*, 38(1), 77-99. <https://doi.org/10.1080/09515089.2024.2310628>
- Kang, S.-P. (2016). Somatoparaphrenia, the body swap illusion, and immunity to error through misidentification. *The Journal of Philosophy*, 113(9), 463-471.
- Kang, S.-P. (2025). Against an epistemic argument for mineness. *Review of Philosophy and Psychology*, 16(1), 47-64. <https://doi.org/10.1007/s13164-024-00723-2>
- Lane, T. (2012). Toward an explanatory framework for mental ownership. *Phenomenology and the Cognitive Sciences*, 11(2), 251-286. <https://doi.org/10.1007/s11097-012-9252-4>
- Liang, C., & Lane, T. (2009). Higher-order thought and pathological self: The case of somatoparaphrenia. *Analysis*, 69(4), 661-668.
- Lichtenberg, G. C. (2012). *Georg Christoph Lichtenberg: Philosophical writings*. State University of New York Press.

- Lotan, G. (2022). Grounding bodily sense of ownership. *Philosophia*, 50(5), 2617-2626. <https://doi.org/10.1007/s11406-022-00554-4>
- Martin, M. G. F. (1995). Bodily awareness: A sense of ownership. In J. L. Bermúdez, T. Marcel, & N. M. Eilan (Eds.), *The body and the self* (pp. 268-289). MIT Press.
- McClelland, T. (2023). Four impediments to the case for mineness. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 50-76). Oxford University Press.
- Nida-Rümelin, N. (2023). Experiencing subjects and so-called mine-ness. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 191-222). Oxford University Press.
- Palmira, M. (2020). Immunity, thought insertion, and the first-person concept. *Philosophical Studies*, 177(12), 3833-3860. <https://doi.org/10.1007/s11098-019-01411-z>
- Peacocke, C. (1999). *Being known*. Oxford University Press.
- Peacocke, C. (2017). Philosophical reflections on the first person, the body, and agency. In Frederique de Vignemont & A. J. T. Alsmith (Eds.), *The subject's matter: Self-consciousness and the body* (pp. 289-310). MIT Press.
- Praetorius, N. (2009). The phenomenological underpinning of the notion of a minimal core self: A psychological perspective. *Consciousness and Cognition*, 18(1), 325-338.
- Prinz, J. (2011). Waiting for the self. In J. L. Liu & J. Perry (Eds.), *Consciousness and the self: New essays* (pp. 123-149). Cambridge University Press.
- Rosenthal, D. M. (2010). Consciousness, the self and bodily location. *Analysis*, 70(2), 270-276.
- Russell, B. (1945). *A history of Western philosophy*. Simon and Schuster.
- Salje, L. (2017). Crossed wires about crossed wires: Somatosensation and immunity to error through misidentification. *Dialectica*, 71(1), 35-56. <https://doi.org/10.1111/1746-8361.12170>
- Salje, L., & Geddes, A. (2023). Conscious experience: what's in it for me? In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 27-49). Oxford University Press.
- Scheur, J. K. (2009). Experience and self-consciousness. *Philosophical Studies*, 144(1), 95-105. <https://doi.org/10.1007/s11098-009-9381-y>
- Serrahima, C. (2023). The bounded body. On the sense of bodily ownership and the experience of space. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 250-268). Oxford University Press.
- Serrahima, C. (2024). Somatosensation and the first person. *Review of Philosophy and Psychology*, 15(1), 51-68. <https://doi.org/10.1007/s13164-022-00660-y>
- Serrahima, C. (2025). Psychological immunity, bodily ownership, and vice versa. *Philosophical Psychology*, 38(3), 1225-1249. <https://doi.org/10.1080/09515089.2024.2412197>
- Shoemaker, S. (1986). Introspection and the self. *Midwest Studies in Philosophy*, 10, 101-120.
- Shoemaker, S. (1968). Self-reference and self-awareness. *The Journal of Philosophy*, 65(19), 555-567.
- Siderits, M. (2011). Buddhist non-self: The no-owner's manual. In S. Gallagher (Ed.), *The Oxford Handbook of the self* (pp. 297-315). Oxford University Press.
- Soldati, G. (2023). Mineness, deflation, and transparency. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 99-119). Oxford University Press.

- Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. The MIT press.
- Wright, C. (2012). Reflections on François Recanati's 'Immunity to error through misidentification: What it is and where it comes from'. In S. Prosser & F. Recanati (Eds.), *Immunity to error through misidentification: New essays*. (pp. 247-280). Cambridge University Press.
- Wu, W. (2023). Mineness and introspective data. In M. García-Carpintero & M. Guillot (Eds.), *Self-experience: Essays on inner awareness* (pp. 120-141). Oxford University Press.
- Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press.
- Zahavi, D., & Kriegel, U. (2015). For-me-ness: What it is and what it is not. In A. Elpidorou & W. Hopp (Eds.), *Philosophy of mind and phenomenology* (pp. 36-53). Routledge.

CARLOTA SERRAHIMA is a teaching and research fellow ('professora ajudant doctora') at the Department of Philosophy of the Faculty of Philosophy and Sciences of Education at the University of Valencia. Her main research interests involve issues related to the philosophy of mind and phenomenology.

ADDRESS: Departament de Filosofia, Facultat de Filosofia i Ciències de l'Educació, Universitat de València. Avinguda Blasco Ibáñez, 30, El Pla del Real, 46010, València, Spain —carlota.serrahima@uv.es—<https://orcid.org/0000-0001-9525-6508>



BERMÚDEZ'S VIEW ON INNER SPEECH: A CRITICAL ASSESSMENT

(La teoría del habla interna de Bermúdez: una evaluación crítica)

Daphne Bernués Moss*

Universitat Pompeu Fabra

<https://orcid.org/0009-0006-7835-832X>

Marta Jorba Grau**

Universitat Pompeu Fabra

<https://orcid.org/0000-0001-6484-2301>

Keywords

Inner speech
Intentional ascent
Conscious thinking
Reflexive thinking
Propositional mindreading

ABSTRACT: José Luis Bermúdez defends the view that inner speech is necessary for thinking about thoughts, what he calls 'intentional ascent'. On his account, we can only take a thought as an object of further thought if the target thought is "held in mind" in inner speech in a way that its canonical structure is revealed. Two paradigm cases exemplify this view: reflexive evaluation and propositional mindreading. In this article, we examine Bermúdez's view and argue that 1) the process of intentional ascent *via* inner speech is empirically untenable and psychologically unrealistic, and that 2) the view, in demanding that words reveal the canonical structure of thoughts, runs into a dilemma where inner speech becomes either superfluous or inaccurately described. We finish by gesturing towards a more encompassing view of the role of inner speech in thinking that, while rejecting Bermúdez's more problematic claims, maintains his main insight of the relevant role of inner speech for human cognition.

Palabras clave

Habla interna
Ascenso intencional
Pensamiento consciente
Pensamiento reflexivo
Mindreading proposicional

RESUMEN: José Luis Bermúdez defiende la idea de que el habla interna es necesaria para pensar sobre pensamientos, lo que él llama «ascenso intencional». Según su teoría, solo podemos tomar un pensamiento como objeto de otro pensamiento si el primero se «mantiene en la mente» en el habla interna de manera que se revele su estructura canónica. Dos casos paradigmáticos ejemplifican esta tesis: la evaluación reflexiva y el *mindreading* proposicional. En este artículo, analizamos la teoría de Bermúdez y argumentamos que 1) el proceso de ascenso intencional a través del habla interna no se sostiene empíricamente y no es realista a nivel psicológico, y que 2) la teoría, al exigir que las palabras revelen la estructura canónica de los pensamientos, se enfrenta a un dilema en el que el habla interna o bien se vuelve superflua, o bien se describe de forma no apropiada. Para acabar, apuntamos hacia una visión más amplia del papel del habla interna en el pensamiento que, aunque rechace las afirmaciones más problemáticas de Bermúdez, mantiene su intuición principal acerca de la relevancia del habla interna en la cognición humana.

Gako-hitzak

Barne-hizketa
Igoera intenzionala
Pentsamendu kontzientea
Pentsamendu erreflexiboa
Gogamen-irakurketa proposizionala

LABURPENA: José Luis Bermúdezek dio barne-hizketa beharrezkoa dela pentsamenduei buruz pentsatzeko; «igoera intenzional» deitzen dio. Haren arabera, pentsamendu bati buruz pentsatzeko, barne-hizketaren bitartez «atxiki behar da gogamenean» pentsamendu hori, haren egitura kanonikoa agerian geratzeko moduan. Bi kasu paradigmatikok argitzen dute ikuspegi hori: ebaluazio erreflexiboa eta gogamen-irakurketa proposizionala. Artikulu honetan, Bermúdezen ikuspegia aztertuko dugu, eta zera argudiatuko: 1) barne-hizketaren *bidezko* igoera intenzionaleko prozesua enpirikoki defendaezina eta psikologikoki sinesgaitza da, eta 2) ikuspegia, hitzek pentsamenduen egitura kanonikoa agerian jartzea eskatzen duenez, dilema baten aurrean kokatzen da, non barne-hizketa ez baita beharrezkoa edo oker deskribatzen baita. Amaitzeko, barne-hizketak pentsamenduen duen eginkizunaren ikuspegi zabalagoa emango dugu: Bermúdezen baieztapen problematikoek baztertzeko baditu ere, barne-hizketak giza kognizioan duen eginkizun garrantzitsuari buruzko haren intuizio nagusiari eusten dio.

* **Correspondence to:** Daphne Bernués Moss. Departament d'Humanitats, Universitat Pompeu Fabra, Barcelona. C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. – daphnececilia.bernues@upf.edu – <https://orcid.org/0009-0006-7835-832X>

** **Correspondence to:** Marta Jorba Grau. Departament d'Humanitats, Universitat Pompeu Fabra, Barcelona. C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain – marta.jorba@upf.edu – <https://orcid.org/0000-0001-6484-2301>

How to cite: Bernués Moss, Daphne; Jorba Grau, Marta (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>).

Received: 13 May, 2024; Final version: 06 March, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2025 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

José Luis Bermúdez has argued in various of his works that inner speech is required for conscious thinking about thinking, that is, for higher-order thought or second-order dynamics. This claim is part of a broader investigation into the role of (natural) language in human and non-human cognition. In his 2003 book *Thinking without words*, Bermúdez develops a framework of thinking in non-linguistic creatures. He argues that non-linguistic thinking reaches a limit in the types of thinking that involve *intentional ascent*, a cognitive process that requires the use of language. His central claim regarding the thought abilities of linguistic creatures is that

[A]ll thinking that involves intentional ascent (roughly, all thinking that involves thinking about thoughts) requires the capacity for semantic ascent (roughly, the capacity to think about words). (Bermúdez, 2003, p. 151)

Thinking about thoughts (intentional ascent) requires thinking about words (semantic ascent), in his formulation. Bermúdez's approach belongs to a family of views called "format views" (Martínez-Manrique and Vicente, 2015), according to which inner speech is the vehicle of thought and the "format" in which some of our thoughts are encoded. Carruthers (2011) and Clark (1998) are also representatives of this approach. Interestingly, Bermúdez (2003, pp. 158-164) argues that Clark's account is correct but incomplete because he shows that we engage in second-order dynamics using language, but for Bermúdez (2003, p. 158) "what we need is an argument that second order cognitive dynamics can only be undertaken by language-using creatures". Hence, his goal is to offer an account of the *necessity* of language (inner speech)¹ in specific forms of human cognition, i.e., second-order thought. To this end, Bermúdez develops and further defends the view with a more detailed argument in the article "Inner Speech, Determinacy, and Thinking Consciously about Thoughts" (Bermúdez, 2018a).

In this paper we aim at engaging with Bermúdez's view on intentional ascent by arguing that the view 1) is empirically untenable and psychologically unrealistic and 2) in demanding that words reveal the canonical structure of thoughts, it runs into a dilemma where inner speech becomes either superfluous or inaccurately described. This discussion leads us to gesture towards a picture in which thinking *with* words occupies the central stage, instead of thinking *about* words. We believe that this alternative avoids certain less compelling claims while remaining congenial to Bermúdez's main insight.

2. The view: *intentional ascent requires semantic ascent*

Bermúdez's view is that there are certain thinking operations that are only available to humans because they are capable of producing language: the operations that involve "explicit" intentional ascent (2003, p. 166) or conscious thinking about thinking. In the 2018a chapter, Bermúdez clarifies the explanandum of the thesis that intentional ascent requires semantic ascent: "[t]he phenomenon under discussion is conscious thinking about thinking —that is to say, consciously taking a thought as an object of thought" (p. 200). This kind of thinking allows us to evaluate our thoughts in accordance with epistemic standards. An example is *reflexive thinking*, or, as stated in Bermúdez (2018a), "reflective evaluation and monitoring of one's own thinking" (p. 201). Reflexive thinking typically includes evaluating evidence for beliefs, such as probable consequences of action, and evaluating inferences between thoughts:

Such reflexive thinking involves having the target thoughts in mind —entertaining them consciously and considering how they relate to each other logically and evidentially. (2003, p. 159)

¹ Bermúdez, actually, defends the necessity of *natural language sentences*. He believes their most common form is inner speech, but grants that one might 'visualize the written form of a sentence' or "think about a thought through a public utterance or inscription" (2018a, p. 203-4).

To evaluate our thoughts in this way, Bermúdez argues, we must apprehend the target (first-order) thoughts consciously and in a way that allows the second-order thought to entertain them (Bermúdez, 2010, 2018a). First-order thoughts (or any thought to which we have access) must be represented a) in a personal-level, conscious vehicle, and b) the vehicle must be linguistic. These requirements exclude that Language of Thought representations can be the vehicles of thought, as they are not consciously available personal-level representations. This vehicle is inner speech, and it is what allows us to perform logical operations between thoughts:

It is clear, moreover, that reflective belief revision will involve explicit consideration of the formal logical relations between thoughts —and we have no understanding of the logical relations between thoughts except when those thoughts have linguistic vehicles. (Bermúdez, 2003, p. 170)

Hence, in his view, the logical structure of a thought becomes conscious and accessible for second-order thought *via* being formulated in inner speech. One cannot think about logical relations between thoughts “without somehow ‘holding in mind’ the thoughts between which they hold” (Bermúdez, 2018a, p. 201). This is why intentional ascent (thinking about thinking) requires semantic ascent (thinking about words). These “words” are inner speech:

What we introspect when we introspect our propositional thoughts in the manner required for the processes of second-order cognitive dynamics is inner speech. (2003, p. 160)

This is Bermúdez's *explanandum*, and the function for which inner speech plays an essential role. In a slightly different formulation: “Rehearsing thoughts in the “perceptible garb” of inner speech enables them to be the objects of further thoughts” (Bermúdez, 2018a, p. 199).

3. *Inner speech and thinking about thinking: an empirically untenable picture*

Firstly, we would say that it is hard to pin down what exactly is meant by “rehearsing thoughts” or “holding them in mind” in inner speech. One interpretation is that “holding thoughts in mind” is a temporally situated step in the second-order process, and that for a second-order thought to occur, one must linguistically articulate the first-order thoughts first in order to grasp them. In another perhaps more charitable interpretation, the first order thoughts “held in mind” are directly embedded or can already be found within the second-order dynamics. In either case, the first-order thought is described as necessarily being “held in mind” in inner speech (i.e., consciously) in the process of thinking about thinking. One possible understanding of what Bermúdez seems to refer to here is that the first-order thought has to be retained in working memory for it to be evaluated, similarly to what happens when we briefly ‘hold in mind’ a phone number to then directly dial it. What this suggests is that by “holding in mind” the linguistically articulated first order thought, he is providing a psychological description of how we engage in second order thinking and the role of inner speech within it.

In Bermúdez (2003), he describes three paradigm cases that require intentional ascent and are *not* available to non-linguistic creatures (see Chapter 9 in Bermúdez, 2003): reflexive thinking, higher-order desires, and propositional mindreading. In Bermúdez (2018), only two of them remain: reflexive thinking (which changes to “reflexive evaluation”) and propositional mindreading. Bermúdez claims that the explicit representations entailed by these processes can only take place in a linguistic, natural language medium. The reason is, again, that we need the ability to “hold thoughts in mind” to think about them. We will argue that there is empirical evidence showing that reflexive evaluation and propositional mindreading can occur without inner speech, and that the way he describes the role of inner speech in intentional ascent also seems introspectively unrealistic.

3.1. REFLECTIVE EVALUATION

The first paradigm case that Bermúdez (2018a, p. 201) presents is reflective evaluation and monitoring of one's thinking:

We engage in such reflective evaluation and monitoring, for example, when we think about the evidential relations between our judgments and the evidence that we have for them. Whether those evidential relations are deductive or inductive, they hold between thoughts or propositions (which for present purposes I take to be equivalent). And so, one cannot think about those evidential relations without somehow “holding in mind” the thoughts between which they hold.

Second-order thought needs inner speech because sustaining thoughts in language in this way allows us to think about the evidential relations that hold between them. This leads to the consequence that any thought from which one consciously derives an inference or draws an evidential relation has to be “held in mind” in language, i.e., inner speech:

[I]t is only possible to evaluate and reflect on the extent to which one belief implies another (...) or the extent to which a belief is supported by a particular type of evidence if one is able *explicitly to hold those beliefs in mind*. (Bermúdez, 2003, p. 170, emphasis added)

This use of inner speech, however, does not find support in the empirical literature. We will argue that there are empirical studies showing that children at the early stage of 4 to 6 years are capable of engaging in conscious reflexive thinking. The relevant aspect to take into account when assessing these experiments is that inner speech is said to develop at the age of 6-7 (Vygotsky 2012/1934; Alderson-Day & Fernyhough, 2015)² which makes it very improbable that children below these ages are using inner speech when they engage in such tasks (more on this below). The empirical literature on forms of reflexive thinking is quite abundant, and developmental psychologists have researched and found evidence for children's abilities of abstract reasoning (Walker *et al.* 2016), discerning causally relevant variables (Goddu *et al.* 2020), revising their higher-order beliefs depending on the strength of the evidence (Kimura *et al.* (2019) and other metacognitive processes such as distinguishing mental states of knowledge from those of uncertainty (Cook *et al.*, 2011). In these studies, however, children's metacognitive abilities are measured indirectly, i.e. inferred from their behavior, and so the studies are evidence for *implicit* forms of reflexive thinking. Such forms would not target Bermúdez's claim, as he specifically focuses on the possibility of engaging in *conscious* or *explicit* reflexive thinking.

There is empirical evidence, however, that shows that children between 4 and 6 years old are also capable of conscious reflexive thinking. A study by Lagenhoff *et al.* (2024) shows that 4- to 6-year-old children seem to think reflexively in this way. In their study, they wanted to know whether exposing children to the disagreement of others would decrease their confidence in their beliefs. First, they taught children a non-existent category (“blicket”). They then asked them to point at those objects that were “blickets” and assessed their confidence in their beliefs with questions such as: “In a scale from 1 to 4: How sure are you that X toys are blickets?” or “Are you really sure or just a little sure?”. They then asked a confederate to point at “blickets” in front of the children. The answers were either in agreement or in disagreement with the children's. When the confederate's answers disagreed with the children's, their confidence judgments decreased significantly in the second measurement.³ Finally, in a subsequent period of “information search”, children in the disagreement condition were found more active. We believe this is an empirical case of children evaluating “the extent to which a belief is supported by a particular type of evidence”, which is one of the possibilities that Bermúdez explicitly considers as reflexive evaluation. Importantly, the confidence questions force children to explicitly reflect on that matter, indicating an *explicit*

² See also Geva & Fernyhough (2019) for an examination of the causal relation between the emerging of inner speech in early childhood and the maturation of the dorsal language stream, and the review of Alderson-Day & Fernyhough (2015) for evidence about the emergence of the phonological effect (usually taken as evidence for covert verbal rehearsal) around 7 years of age.

³ Children's confidence decreased *only* when the confederate was introduced as an expert, which, according to the authors, shows children's sensitivity to the epistemic status of whom they were disagreeing with.

metacognitive capacity (and thus, distinguishing this experiment from the other results on implicit reflexive thinking). One possible source of skepticism with this case is that perhaps older children in the group, i.e., those 6 years old, could be using inner speech for assessing the confidence in their beliefs. However, no significant differences were found in the result between 4-5 and 6-year old children, which seems to discard this possibility.

Another relevant study without this potential worry related to age is the one by Gardier & Geurten (2024), where they carefully measured the metacognitive abilities of children between the ages of 2.5 and 4.5. In this study, infants were presented with a memory recognition task. After each forced-choice memory decision, they measured their “retrospective confidence judgment”. This measure was two-fold. First, they asked them to indicate their degree of certainty using two pictures, i.e., pointing at a picture with a confident character if they were sure of their answer, or another one with a doubtful character if they were not. Second, for each trial, children were given the possibility of asking for a cue when they thought they had given an incorrect answer. They were told by the experimenters that “they had to ask for the cue only when they felt like they needed it to select the correct response” (p. 1248). In both studies, then, under the term “metacognition”, children are tested to see whether they were able to think about their own thinking, e.g., to judge their confidence in their beliefs based on their observations and weighting the beliefs of others. The explicit nature of the task (i.e., reporting and considering the need for help) ensures that children are having a *conscious* thought about their thoughts, and that the results are not explainable, for instance, by there being subpersonal mechanisms subconsciously “tipping the scale” in their decisions.

These two studies presented show cases of conscious reflexive thinking without inner speech, thus putting pressure on Bermúdez' account. However, it is true that such studies *do not rule out* the possibility of children engaging in private speech, that is, talking out loud, while they answer the confidence questions. Private speech, in this case, would presumably allow them to “objectify their thoughts” so that they could assess the evidence they have for them. Note, however, that not any form of private speech would help Bermúdez's case, only the hypothetical case in which children in the experiments utter and “hold in speech” the thought that is the object of metacognition (more discussion on this in Section 4).

This potential problem is not found in other studies that examine inner speech in reasoning using articulatory suppression techniques (Nedergaard *et al.*, 2023), thus being more conclusive evidence against Bermúdez's account. In articulatory suppression, the ability to rehearse verbal information is blocked by instructing subjects to perform a task while producing language out loud. If the performance of the task is affected, it is concluded that the task had a linguistic component. In the study of Evans & Brooks (1981), they tested whether phonological suppression had an effect on revising the logical validity of inferences. In their experiment, subjects were presented with various valid or invalid logical inferences (e.g., If it is not a triangle, then it is not red. It is not red. Then, it is not a triangle) and they had to decide whether or not the conclusion follows logically. To impede linguistic rehearsal for solving the problems, a group of subjects had to utter a counting sequence repeatedly while solving the task (the “articulation condition”). Overall, this experiment required that subjects engaged in reflexive evaluation and thought about the (deductive, in this case) “evidential relations between our judgments and the evidence that we have for them” (to use Bermúdez's phrasing). Not being able to use inner speech would impair the reasoning, according to him. However, the results showed that subjects in the articulation condition did not perform any worse (surprisingly, they were found to answer faster than the control group).

While we believe such empirical evidence is conclusive regarding the necessity of inner speech for intentional ascent in the case of reflexive thinking, we also think that it is worth pointing at some *daily-life* situations in order to intuitively doubt the idea that thinking about thinking necessitates inner speech as the format that “holds in mind” our thoughts. Bermúdez does not give the reader any examples of how the reflective process presumably unfolds and how inner speech plays its role (which makes this rather intuitive assessment of the theory a bit harder), but we contend that daily-life situations that are potential candidates of reflexive thinking seem to occur without “holding in mind” in inner speech the thoughts we are thinking about.

SCHOOL HOLIDAY

I observe from the street that the light in the apartment I share with my partner is on, and it is a moment of surprise because, at this time, I do not expect her to be home from teaching at the school. While I look for reasons why this would be so, I remember seeing an image of a newspaper front page that said that today is a city holiday. Then, I remember her telling me that one of the perks of this job was that, whenever there is a city holiday, schools end classes earlier than normal. I conclude, then, that my partner is home, and that's why the light is on.

Answering the question of why my apartment's light is on entails reflection on "the extent to which one belief implies another" (Bermúdez, 2003, p. 170). The kind of reasoning going on in this case is deductive reasoning whose premises could be something like: "It's a holiday; if it's a holiday, school classes end at midday and my partner will be home early; school classes end at midday and my partner will be home early". However, each premise of this deduction need not be enunciated in inner speech; it suffices to remember the newspaper front page from earlier in the morning by which you learned about the holiday to reach the conclusion that my partner is home, which might or might not be enunciated in inner speech. Notice that we are not presenting a case of unconscious reasoning, but rather how an episode of reflective thinking would unfold consciously. We believe that this ordinary example, and other similar ones, intuitively motivate the idea that this sort of reflexive thinking can be done ordinarily without necessarily using inner speech and, were our reflective operations in need of uttering every proposition in inner speech, this would be too cognitively costly.

A couple of remarks are in order before moving to the case of propositional mindreading. First, we acknowledge that this intuitive example might not alone convince Bermúdez or anyone favoring his account, but we think that it can, at least conditionally, add more pressure on his account on top of the empirical cases presented. Second, in presenting these empirical and intuitive cases against the thesis of intentional ascent, we are not denying that inner speech might enhance, assist, and be used in many cases of conscious reflexive thinking and evaluation, mostly in adults and children after 6-8 years old. Its cognitive role in *supporting* and *enhancing* reasoning processes, as well as other executive functions such as inhibition, task-switching and planning is well documented in inner speech studies (Alderson-Day & Fernyhough, 2015; Petrolini *et al.*, 2020), which overall speaks in favor of weakening the claim of the necessity of inner speech for such cognitive processes.

3.2. PROPOSITIONAL MINDREADING

The second paradigm case for intentional ascent is propositional attitude mindreading, that is, "consciously attributing thoughts to others in order to make sense of and predict their behaviour" (Bermúdez 2018a, p. 201). The kind of propositional attitude mindreading Bermúdez focuses on is to be distinguished from other forms of social coordination in which no explicit attribution of thoughts is done, and importantly, from other forms of mindreading such as perceptual mindreading, in which subjects simply "need to be able to represent the fact that the agent is perceiving that state of affairs" (Bermúdez 2018a, p. 202). Whereas perceptual mindreading is world-oriented, and attributes a state of affairs as perceived by another subject, propositional mindreading involves metarepresentation, and entails representing how the subject is *representing* a state of affairs (propositions, thoughts, or belief-states held by the subject) (Bermúdez, 2003; 2018b). Also, propositional states of others must be consciously accessible to consider them in one's course of action; they are conscious insofar they are a part of "conscious practical decision-making" (2018b, p. 124). Conscious propositional mindreading is not possible without inner speech, again, mediating metarepresentation or the process of "holding a thought in mind". In Bermúdez's phrasing: "it is impossible explicitly to attribute a propositional attitude to another subject without explicitly representing the thought that is the content of the propositional attitude" (2018b, p. 202).

The problem, as with the previous case, is that there is empirical evidence against conscious propositional mindreading necessitating inner speech. Propositional mindreading is standardly measured with false-belief paradigms (Wimmer & Perner, 1983), and false-belief tasks are considered by Bermúdez (2018b) cases of propositional mindreading. There has

been much discussion regarding the role of language in belief attribution, both in children and adults (Milligan *et al.*, 2007; Newton and de Villiers, 2007, among others), but here we will focus on Forgeot d'Arc & Ramus (2011), who present a study with people from ages 19-31 to specifically examine the role of inner speech in belief attribution. For this, they use a completely nonverbal paradigm consisting of silent animated cartoons. Subjects were presented an animated film until the suspense scene froze, and they had to choose the most plausible end of the story from the two that appeared on the screen and give their response on a response box. The study used verbal shadowing in the dual-task paradigm to inhibit inner speech, so that "a sufficiently demanding verbal shadowing task engages phonological representations to such an extent that it is not possible to concurrently formulate propositions using inner speech" (Forgeot d'Arc & Ramus 2011, p. 978). The concurrent shadowing task consisted of repeating each word as it was being heard from their headphones. Results show that verbal interference had an effect on performance, which decreased, but no effect on belief attribution. Participants were able to attribute beliefs despite having their inner speech interfered with, and hence, their results spoke against the hypothesis that verbal rehearsal is necessary for belief attribution. The attribution of the belief is *conscious* because participants were explicitly asked for the end of the story.

In the study of Dungan & Saxe (2012), they used a similar procedure of showing participants a video of live actors who were dressed as animals, and in each video a rabbit hid carrots under a box, then a cat came and moved the object first to a second box, and then to a third box. During the movements, the rabbit was either present ("true belief") or absent ("false-belief"). Participants had to answer for the most likely ending of the video. The study tested the effects of different kinds of shadowing (verbal shadowing, no shadowing, or rhythms) and found that performance hindrance was not specific to the verbal kind. Overall, then, there seems to be empirical evidence from these different domains suggesting that conscious belief attribution functions quite independently from inner speech.

As we also argued for the case of reflexive evaluation, we contend that the use of inner speech described by Bermúdez seems psychologically inaccurate in daily-life situations. Imagine the following case of propositional attitude attribution:

A TEENAGER IS FOUND OUT

A teenager wants to go out with her friends when she is not supposed to. She says she's going to sleep and hides pillows under her blanket to form the impression that she's sleeping in her bed.
The teenager sneaks out, but her mom finds out.

In Bermúdez's account, the parent would need to 'hold in mind' and explicitly represent (using inner speech) the thoughts that she will attribute to her daughter. For instance, that she wanted to pretend she was sleeping; that she thought that the pillows would do the trick, and that she does not yet know that she has been caught. (The last one is a false-belief attribution: the mom attributes to the daughter the *false belief* that the plan worked.) However, the mom can propositionally mindread her daughter in various ways without having to hold in mind or enunciate such thoughts in inner speech. The propositional states of her daughter are consciously accessible to her and can take part in her conscious, practical decision-making' (Bermúdez, 2018b). For instance, she might decide to amuse herself the next day and ask her trick questions about how she could possibly get any sleep given the unusual noises coming from the street, and she might mindread her further and predict that she will probably evade these questions by making more things up. So, the mother can attribute thoughts to her daughter, plan to act accordingly (ask her tricky questions the next day) and predict her behavior (creativity in her answers), without having to formulate and consciously access a linguistic representation of her daughter's belief and without articulating anything in inner speech. It should be noted that of course it is possible to reconstruct the case by making explicit in inner speech each of the steps in the process of belief attribution but this doesn't seem the typical way of proceeding.

3.3. INTENTIONAL ASCENT IN AN EPISTEMIC READING

A possible reply to our criticisms is that Bermúdez is not trying to describe the *psychological* process by which these second-order operations happen but rather he would be taking an *epistemic* approach, addressing the question of what

would confer justification and internal epistemic warrant to our reasoning process: what justifies me in believing that the inference I draw is valid?⁴ The idea here would be that inner speech is necessary for this internal justification-conferring process —what we would have to answer if asked for justification (i.e., language, sentences, *words*). Under this epistemic reading, Bermúdez could agree that psychologically inner speech is not required for intentional ascent and propositional mindreading and still maintain that explicit inner speech representation is epistemically necessary for intentional ascent, that is, needs to be invoked in any process of conferring justification to the deliberative processes leading to conclusions (and, plausibly, to attributing propositional states to others).

One should notice, however, that this interpretation is not straightforward in Bermúdez's works. In many places, he is clearly speaking about a psychological process, while in others, he only mentions epistemic notions such as 'evidence' or 'evaluation' (Bermúdez, 2018a). In the most recent argument for intentional ascent (see Section 4 below), specifically, the epistemic interpretation doesn't seem warranted. But even if we grant this epistemic interpretation to avoid the worries presented and others, additional problems arise for this reading, as one could doubt that inner speech is required for such an epistemic process of reasoning. That is, would higher-order thinking *just* be epistemically justified when carried out in inner speech? The kind of epistemic justification at stake here has to be *internal* justification, for clearly one cannot defend that the validity of the deduction (something external) depends on my rehearsing in inner speech the premises and the conclusion (*pace* certain forms of *psychologism*). The problem for internal justification appears, though, for the cases of inner speech absence, as the ones presented above for reflexive evaluation and propositional mindreading (and see also cases of unsymbolized thinking or anendophasia presented below). It seems reasonable to say that subjects in these cases can be epistemically justified without relying on inner speech.

4. *A dilemma for the role of inner speech in intentional ascent*

The most detailed argument Bermúdez offers for his thesis that intentional ascent requires semantic ascent can be found in his 2018a article (pp. 202–203). In this Section, we will present the argument, comment on and add existing evidence against the claim that the vehicles of the intentional ascent must be either language-like or image-like (Premise 3), and focus on a key idea of the argument (Premise 4): target thoughts must be represented in a way that reveals their canonical structure (i.e., how they are built up from their constituent elements). We argue that the way Bermúdez believes the canonical structure of natural language sentences might be settled, that is, subconsciously, is in tension with the idea that expanded forms of inner speech are better suited for the job of revealing it, leading to a dilemma for the view.

Bermúdez's (2018, pp. 202–203)⁵ argument runs as follows:

1. The paradigm cases of thinking about thinking are conscious mental acts that involve taking a target thought as the object of thought.
2. So, these target thoughts must be represented by consciously accessible constituents of a creature's psychological life.
3. The representational format for these consciously accessible constituents of psychological life must be either language-like or image-like.
4. Target thoughts must be represented in a way that reveals their canonical structure (i.e., how they are built up from their constituent elements).
5. Representing thoughts in a pictorial or image-like format does not reveal their canonical structure.

⁴ We are indebted to Zoe Drayson (UC Davis) for calling the attention to this reading in the talk "Inner Speech and Thinking about Thoughts", delivered on March 14, 2023, within the Inner Speech Seminar Series organized by Daniel Gregory with the collaboration of the research project "Inner Speech in Action" (UPF), the LOGOS Research Group and the Barcelona Institute of Analytic Philosophy.

⁵ As mentioned above, notice that in this detailed argument there is no trace of the 'epistemic reading' present, but rather what is at stake is the characterization of the psychological process of second-order thinking and what sort of representational format is required.

6. The canonical structure of a thought is revealed when it is represented in a linguistic format.
7. Thoughts can be linguistically represented either by sentences in the language of thought, or by natural language sentences.
8. Sentences in the language of thought are not consciously accessible constituents of psychological life.
9. Therefore, thoughts can only be taken as the objects of thought when they are represented by natural language sentences.

First, let us refer briefly to the possible problems posed by Premise 3: “The representational format for these consciously accessible constituents of psychological life must be either language-like or image-like.” As other authors have already highlighted (Lurz, 2007), this premise would deny that in cases of unsymbolized thinking (Hurlburt and Akther, 2008), intentional ascent is possible. Unsymbolized thinking is an episode of thinking without any words or images that some people report having as an inner experience, characterized by Hurlburt and colleagues. It seems difficult to say that in those cases intentional ascent is not possible, that is, that an unsymbolized thought cannot become the object of a second-order thought. Minimally, saying ‘this thought’, that is, ostensibly referring to an unsymbolized thought (which is the kind of thing subjects might report in interviews with Hurlburt and colleagues), is a second-order thought whose target thought is not linguistically articulated, and therefore its canonical structure cannot be revealed (Premise 6)⁶. Moreover, in the interviews by Hurlburt and colleagues, subjects are asked to describe the experience of the thought in detail, thus resulting in not just its apprehension but a thorough examination and reflection of the thought itself.

Analogously, cases of anendophasia (Nedergaard & Lupyan, 2023), in which subjects report experiencing no inner speech at all, would present an additional problem for premise 3. Are people with anendophasia incapable of intentional ascent? This is simply not the case. Rather, their behavioral differences are seemingly minimal (e.g., in Nedergaard & Lupyan (2023), poorer performance in rhyme judgment tasks and verbal working memory) and *not* crucial to their cognition, as Bermúdez’s account would predict. Anendophasia is a phenomenon discovered only recently, and the empirical literature on its psychological implications is nascent. We would argue that the fact that this condition has remained unnoticed for so long speaks against the necessity of inner speech (insofar as it denotes a conscious, introspectable phenomenon) for any process that would translate into crucial behavioral differences, e.g., not being able to think reflexively, or the impossibility of attributing propositional states to others.

In the case of anendophasic individuals, one could argue that perhaps they could be talking out loud or writing in order to perform tasks related to intentional ascent. While it is true that private speech seems to substitute inner speech usage for anendophasics (for instance, the study of Nedergaard & Lupyan (2023) found that performance in the verbal working memory and rhyme judgment tasks decreased only in anendophasics who *did not* talk out loud while performing them), no studies have proven this compensation with regards to reflexive evaluation and propositional mindreading. And while it is possible to imagine how anendophasics could particularly benefit from uttering out loud or writing their thoughts for reflexive evaluation, the fact that they would have to do the same with the propositional states of others in order to attribute them seems far-fetched. Lastly, both in unsymbolized thinking and anendophasia cases, Bermúdez could appeal to a recharacterization of the cases as misdescriptions, illusory experiences, or as somehow involving some minimal form of inner speech, despite what subjects report. However, this comes at the cost of denying the phenomena *per se*. Accepting instead that second order thought is not possible for these subjects also seems a high cost to pay.

⁶ We are also inclined to doubt Premise 5: thinking about thinking can only be done in a linguistic format (see also Munroe, 2021, for a defense of image-based reasoning). Image-like formats, such as following a pattern of visually representing different premises of an argument seem to be able to do the job. In a reasoning of the sort ‘Jan is taller than Júlia’, ‘Júlia is taller than Mía’, therefore ‘Jan is taller than Mía’, we might not hold in mind these sentences in inner speech but just visually represent what they express in order to evaluate their correction. We won’t pursue this debate here, but see Lurz (2007), and also Bermúdez (2011, 2018b) for replies.

Now let's move to Premise 4 of the argument. It states that for intentional ascent we need a representational format that reveals the canonical structure of thought, that is, "how [thoughts] are built up from their constituents". A first point to note is that, as several authors have emphasized (Vygotsky 2012/1934; Fernyhough 2004), inner speech normally comes in a *condensed* format, that is, as fragmented words such as 'supermarket'. Such "single words" are not apt to reveal the constituents of thoughts. Indeed, Bermúdez admits that "Beckett-like subsentential utterances *are typically not useful tools* for bringing full-fledged propositional thoughts to mind" (Bermúdez 2018a, p. 205, our emphasis). Thus, there needs to be a certain inner speech profile that allows for the revelation of canonical structure, namely, *expanded* forms of inner speech (*ibid.*).

However, Martínez-Manrique and Vicente (2010) have already objected that inner speech cannot encode thoughts because of its semantic indeterminacy, that is, natural language sentences do not determine the propositions by themselves, and so target thoughts cannot become the object of further thought *via* inner speech. Bermúdez (2018a) responds in detail to this objection. First, he recognizes that context must play a role, and is tempted to reject the semantic indeterminacy of inner speech on the basis of the crucial role of context in the determination of meaning, but he decides not to pursue this strategy. In the end, he grants natural language's indeterminacy and offers a way to avoid their criticism while maintaining his theory: semantic and syntactic determination can be solved by *subconscious operations*, and what is apprehended in inner speech is the determinate content of the sentence itself. In inner speech, then, what disambiguates and fully determines its meaning are subpersonal computations, leaving it open whether indeterminacy is experienced or not at the conscious level (p. 211).

We believe that this response leaves Bermúdez's account in a problematic situation regarding the role of inner speech in presumably revealing the canonical structure of our thoughts (Premise 4). On the one hand, if content is determined subconsciously, then highly condensed inner speech (and arguably, "Beckett-like subsentential utterances") *would do equally well* in revealing the canonical structure of the thought in play. But in how Bermúdez initially sets the stage, these utterances are "typically not useful tools for bringing full-fledged propositional thoughts to mind" (p. 205), expanded kinds being the ones suited for the job. The fact that the degree of expansion of inner speech has any influence in revealing the canonical structure becomes something hard to explain once it is assumed that syntactic/semantic ambiguity is solved subconsciously. Why would more words reveal it better? Intentional ascent would be equally feasible over "I need to go to the supermarket before it closes" than the same-meaning utterance "supermarket!". Thus, subconscious determination of content renders inner speech expansion superfluous.⁷

Here Bermúdez could argue that we are conflating 'content determination' with 'revelation of canonical structure', so that his view is in fact that while content is determined subconsciously, inner speech is required for *revealing* thought's canonical structure consciously. In this case, inner speech would still be key in making *conscious* the structure and constituents of a first-order thought, which would be something "added" to the determination of content, and which could plausibly be dependent on the properties (e.g., the expansion) of the inner speech vehicle. Bermúdez, however, gives no hint of this dissociation of the expressions 'content determination' and 'revelation of canonical structure' in his answer to Martínez-Manrique and Vicente's (2010) criticism. (If these expressions didn't conflate, Bermúdez's answer would not be an answer to their critique). In any case, we would say that we are sympathetic to the idea that inner speech is (one way of) making thought conscious in general, although in a way that departs from Bermúdez's original view of inner speech being required as stated in Premise 4 (see Section 5).

On the other hand, if one follows Bermúdez's claim that expanded inner speech utterances are better suited to do the job of revealing the canonical structure, we run into the second horn of the dilemma. Indeed, if it is *about words* that we think when we think about thoughts, then the amount of words (expansion) would seem decisive in the revelation of the canonical structure of a thought. The problem is that it is unclear how much expansion is required in order for a

⁷ And if one takes unsymbolized thought and anendophasia cases at face value, then —for the job described by Bermúdez— the role of inner speech alone would be superfluous too.

thought to “be represented in a way that reveals their canonical structure (i.e., how they are built up from their constituent elements)”. Similar to Martínez-Manrique & Vicente’s (2010) criticism of ambiguity, one could insist that even very expanded forms of inner speech might not adequately reveal a thought in the way that Bermúdez seems to require. Indeed, it was Frege’s (1892) and Russell’s (1905) idea (and many others after them) that natural language misleads us about the nature of the propositions we assert, and so of our thoughts. A *superficial* inner speech utterance might lead to a very different proposition once there has been revelation of its *logical form*. The logical form of propositions can be made manifest by the structure of a sentence in a formal language. This format would indeed guarantee the revelation of how a thought is built up from its constituent elements. Such formal languages, though, are not natural public languages like inner speech, for which we would need to “say” in inner speech the utterances that reflect the logical structure of propositions. This scenario would require “holding in mind” complicated sentences (e.g., for every ‘x’, if ‘x’ has the property ‘y’...) and a good number of them (plausibly all the reasoning steps involved). The idea that inner speech must occur in this form is a *reductio ad absurdum* of Bermúdez’s argument. It may only correspond to philosophers and logicians, and in any case, it is not the sort of inner speech profile we typically experience, even when second-order reasoning. In fact, in analogy with cognitive operations such as counting, complex cognitive linguistic activities might normally need the support of the act of writing the sentences down on a paper, a blackboard, etc., as the embodied and extended cognition tradition has highlighted (Bruineberg & van den Herik, 2021). If sophisticated higher-order cases of thinking require extended cognition, then the operation of ‘holding in mind’ might not be sufficient to start with. We have acknowledged earlier, though, that Bermúdez recognizes that the language object of intentional ascent might be written. But notice that if for the most paradigmatic intentional ascent operations we need external support, then his theory might not be best described as a theory of *inner speech*. The fact that natural language sentences (in inner speech, overt speech, or written language) may be beneficial or particularly useful for intentional ascent is not what we are objecting to here. Rather, it is the claim that we would *need* to enunciate our thoughts in a (conscious) natural language (hence, typically inner speech) to think about them.

Notice that our argument resembles Martínez-Manrique and Vicente’s (2010) ambiguity argument, in that the conclusion is the same —inner speech is not suitable for intentional ascent in the way required by Bermúdez— but takes another route to this conclusion: if we grant that expansion is relevant to the canonical revelation of content, then the extreme secure form in which we know that a sentence reveals its canonical structure (i.e., how it’s built up from its constituent elements) is by formalizing it, thus distancing from how typically inner speech is presented in experience. But if such formalization is “said” in words internally, then the process might become too complex to be able to “hold it in mind”, which raises doubts about whether inner speech is the correct *explanans*.

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. In any case, an explanation is lacking on where to draw the line on the expansion needed, and the problem (and hence, the dilemma) is that, as long as the line is not drawn, inner speech seems either trivial or inaccurately described.⁸

Finally, another point to be made is that, while there might be reasons to argue for expanded and logically rich forms of inner speech for the first paradigm case, reflexive evaluation and monitoring of one’s own thinking, one could doubt that the case is applicable to cases of propositional mindreading. While, following Bermúdez, both cognitive activities require metarepresentation, it seems that only the former can insist on the high standard of needing the canonical structure revealed, i.e., for further operations, like revision of evidential relations, contrasting hypotheses, and so on.

⁸ For an alternative view according to which conscious thoughts are determined by the linguistic mechanism underlying inner speech and unsymbolized thought, see Vicente & Jorba (2019).

For the act of consciously attributing a propositional state, however, focusing on inner speech as the format that needs to reveal the canonical structure of the attributed thought (at the cost of being always unrealistically expanded, for instance) remains unjustified.

5. *Gesturing towards a less demanding and more encompassing role of inner speech*

After assessing Bermúdez's theory of intentional ascent, we believe that its high standards might come with a certain "tension" with what seemed to be the project (and the letter) of the 2003 book, namely, the idea of language as mediating the *distinctiveness* of human cognition through the capacity for metarepresentation or metacognition. The uneasiness would come from noticing that what seemed to be a very pervasive and relevant phenomenon in our cognitive lives (that regarded a seemingly "looser" notion of reflexive thinking and included second-order desires) risks being a very infrequent and sophisticated psychological —if not philosophical— process, arguably only available to those logically skilled instead of the linguistically able. A perhaps more congenial picture to his own project as stated in the 2003 book, would construe intentional ascent in a *less cognitively demanding* way, so that *then* it could be seen as more pervasive in our cognitive lives.

We believe that we can maintain the main insight of Bermúdez's views without endorsing his more (to our minds) problematic claims, and unrealistic psychological picture of intentional ascent. It seems plausible that thinking about thinking requires metarepresentation, but not necessarily *via* inner speech. We do not deny that inner speech might sometimes present an extended profile, which may do better at revealing the constituents of thought, and which one may conveniently think "about" in order to think further, but we have argued that to maintain that this is always the case is to demand too much from second-order operations, and from inner speech, which usually presents itself in a more condensed manner, or it is altogether absent, as in the empirical cases presented.

Inner speech can be, among other things, the vehicle in which certain thoughts are embedded, an operation involved in making them conscious. In this sense, we might maintain the importance of inner speech in *making thoughts conscious*, that is, in bringing thoughts to consciousness, in gaining access to our thoughts. This facilitates thought (first and second-order) in a way that creatures who don't enjoy inner speech might lack.

Bermúdez (2018a) raises the point that there is an equivocation between two different senses of 'consciously accessible' that leads to a misunderstanding of his view. The two senses are 1) a thought being consciously accessible insofar as it is conscious and 2) a thought being consciously accessible insofar as it is being consciously thought about (i.e., insofar as it is the object of a conscious thought). He warns us of the risk of conflating both senses. While he accepts that inner speech is not required for 1), inner speech is required for 2), according to him. While we agree with this conceptual distinction, we believe there is room for arguing that in both cases, inner speech can play the same role of "bringing thoughts to consciousness", that is, of making thoughts consciously accessible in the linguistic format, and, importantly, in neither case would inner speech be necessary. If inner speech is the vehicle of some thoughts, understood as the "clothing" in which some thoughts are articulated, this role is done both in 1) and in 2), and noticing this fact might allow us to see a more pervasive role of inner speech in our minds —without endorsing the problematic necessity claim of inner speech, which Bermúdez only holds for 2).

However, we do not need to commit to this being the *only way* in which thoughts come to consciousness and are thought about, as we can also think about pure contents, as in cases of unsymbolized thinking (Hurlburt & Akther, 2008; see Jorba & Vicente 2014 for a defense of such a view). Yet, we can consider language as a privileged format to make our thinking processes conscious and explicit (e.g., when asked for justification, or in specially sophisticated cases), without postulating its use in the manner described by Bermúdez, nor restricting second-order abilities to infrequent and sophisticated processes that, arguably, may not be available solely by being language-using creatures. On the other hand, lowering the demands on inner speech allows one to include many other ways in which inner speech can aid cognition.

If we focus on how inner speech *gives rise* to thoughts, as it occurs in cases of first-order linguistic thinking, and consider it analogous to second-order thinking, we gain a way of rescuing the relevance of inner speech for cognition, in line with Bermúdez's argumentation in the 2003 book. The picture of the necessity of inner speech just for paradigmatically very sophisticated cases of intentional ascent risks jeopardizing Bermúdez's (right) intuition of the relevance of inner speech, and of language altogether for cognition. Perhaps this risk is not real if we drop the phrasing of "thinking *about* words" (which, we argued, is problematic). When we explore our ability (instead of our necessity) to think *with* words, words retain their benefit as a cognitive tool and their ability to bring thoughts (and thoughts for further thoughts) to the forefront, and conduct thought processes consciously. It is in this sense that Jorba & Vicente (2014, p. 95) say:

[W]hen we focus our attention on something by means of inner speech, we become conscious, not of our own inner speech—which is already a conscious phenomenon—but of the *content* of our inner speech. That is, we become conscious, e.g. that the triangle should not be put in the upper left corner, that *p* follows from *q*, or whatever. Inner speech, by driving our attention to contents, makes these contents conscious. And note that inner speech is used in having conscious thoughts, not in having thoughts about those thoughts.

In this picture, inner speech needs not be recruited as a format in order to engage in second-order thinking, and first-order thoughts need not undergo a linguistic transformation to be part of the process. Inner speech might come into play, but not to 'hold in mind' the premises and be the 'blackboard' for evidential relations. It might express, for instance, the end of a second-order process directly (in the SCHOOL HOLIDAY example, the person could say "Ah, she's home early"). It can also take place in "impoverished", condensed inner speech, because in this case it is mainly and directly an operation of the syntactic/semantic aspects of inner speech; a case of thinking about *thoughts*, not about *words*. In this picture, inner speech is a tool that we think *with*, not *about*. Thinking with words is yet one use among others of inner speech, and even if it brings a number of benefits, it might not be necessary for thinking about thinking, if our criticisms and others are on the right spot. If we conceive of inner speech as an *activity* of speaking, albeit silently, in analogy with outer speech (Jorba & Vicente, 2014; Martínez-Manrique & Vicente, 2015), and understand processes of thinking (from the most basic to the most sophisticated) as "piggyback riding" such activity of speaking, we might have a more handy explanation of the relevance of inner speech in our conscious and reflective behavior and cognition. Interestingly enough, this less restrained view (thinking *with* words) seems to represent more accurately the human counterpart mode of thinking implicitly stated in the title of his 2003 book, *Thinking without words*.

6. Concluding remarks

We have presented Bermúdez's view on intentional ascent and the role of inner speech in it, arguing that the way Bermúdez describes second-order thinking processes (those that require intentional ascent, hence semantic ascent, and hence, inner speech) is empirically untenable, and unrealistic when applied to real-life examples. We have also argued that requiring that the target thought be represented and apprehended in inner speech in a way that reveals its canonical structure (Premise 4 of the argument), results in a dilemma where inner speech becomes either superfluous or inaccurately described. Inner speech can be condensed, or it can be absent altogether, and second-order dynamics might still take place (see the empirical evidence above and the intuitive counterexamples). Given its high standards, Bermúdez's view risks depicting inner speech in a too sophisticated form, present in activities that may not be available to the linguistically able, but to the logically skilled, and such activities may need the aid of external support (e.g., writing), thus raising doubts about Bermúdez's view being about inner speech. The thinking processes that require intentional ascent (reflexive thinking and propositional mindreading) are prevalent, widespread, and take place in less demanding settings. Ideally, they are part of the average-human cognitive toolkit. We suggested that a way to account for this fact and to safeguard Bermúdez's role of language in cognition is to reconsider our ability to think *with* words (instead of *about* words), in a way that is more psychologically realistic and less cognitively costly for thinking about thinking to occur,

and for inner speech to be involved. Language can retain the benefits and thought-enabling characteristics defended by Bermúdez, while addressing the distinctiveness of human cognition in a wider variety of cognitive processes and avoiding the necessity claim that, under scrutiny, appears problematic and/or unrealistic.

Acknowledgements

We want to thank Agustín Vicente and two anonymous reviewers for their helpful comments on an earlier version of this paper. This article has been financially supported by the Ministerio de Ciencia, Innovación y Universidades (MICIU) and the Agencia Estatal de Investigación (AEI), research project “Inner Speech in Action: New Perspectives” (INACT) (Ref: PID2020-115052 GAI00/ MICIU/AEI/10.13039/501100011033). It has also received support from the Fundación Ramón Areces, research project “Communication, Inner Speech and Mental Health” (CIM) and from the LOGOS Research Group in Analytic Philosophy (2021-SGR-00276).

REFERENCES

- Alderson-Day, B. & Fernyhough, C. (2015). Inner speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological bulletin*, 141(5), 931-965. <https://doi.org/10.1037/bul0000021>
- Bermúdez, J. L. (2003). *Thinking without words*. Oxford University Press.
- Bermúdez, J. L. (2010). Two Arguments for the Language-Dependence of Conceptual Thought. *Grazer Philosophische Studien*, 81(1), 37-54. https://doi.org/10.1163/9789042030190_004
- Bermúdez, J. L. (2011). Mindreading and moral significance in nonhuman animals. In T. L. Beauchamp & R. G. Frey (eds.). *Oxford handbook of animal ethics* (pp. 407-40). Oxford University Press.
- Bermúdez, J. L. (2018a) Inner Speech, Determinacy and Thinking Consciously about Thoughts. In Langland-Hassan, P. and Vicente, A. (eds.) *Inner speech: new voices* (pp. 199-220). Oxford University Press. <https://doi.org/10.1093/oso/9780198796640.003.0008>
- Bermúdez, J. L. (2018b). Can nonlinguistic animals think about thinking (pp. 119-130). In K. Andres Andrews and J. Beck (eds.). *The Routledge handbook to the philosophy of animal minds*. London: Routledge.
- Bruineberg, J. P., & van den Herik, J. C. (2021). Embodying mental affordances. *Inquiry*, 1-21. <https://doi.org/10.1080/0020174X.2021.1987316>
- Carruthers, P. (2011) *The opacity of mind: an integrative theory of self-knowledge*. Oxford University Press.
- Cook C., Goodman N. D. & Schulz L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 120(3), 341-9. <https://doi.org/10.1016/j.cognition.2011.03.003>
- Clark, A. (1998) Magic words: How language augments human computation. In P. Carruthers and J. Boucher, (eds.) *Language and thought: interdisciplinary themes* (pp. 162-183). Cambridge University Press. <https://doi.org/10.1017/CBO9780511597909.011>
- Dungan J. & Saxe, R. (2012). Matched False-Belief Performance During Verbal and Nonverbal Interference. *Cognitive science*, 36(6), 1148-1156. <https://doi.org/10.1111/j.1551-6709.2012.01248.x>
- Evans, J.S.B.T. & Brooks, P.G. (1981). Competing with reasoning: A test of the working memory hypothesis. *Current Psychological Research* 1, 139-147. <https://doi.org/10.1007/BF02684486>

- Fernyhough, C. (2004). Alien voices and inner dialogue: towards a developmental account of auditory verbal hallucinations. *New ideas in Psychology*, 22(1), 49-68. <https://doi.org/10.1016/j.newideapsych.2004.09.001>
- Frege, G. (1892). "On Sinn and Bedeutung", reprinted in Beaney, M. (ed). (1997) *The Frege reader* (pp. 151-180). Blackwell Publishing.
- Forgeot d'Arc, B., & Ramus, F. (2011). Belief attribution despite verbal interference. *Quarterly Journal of Experimental Psychology*, 64(5), 975-990. <https://doi.org/10.1080/17470218.2010.524413>
- Gardier, M. & Geurten M. (2024) The developmental path of metacognition from toddlerhood to early childhood and its influence on later memory performance. *Developmental Psychology*, 60(7), 1244-1254. <https://doi.org/10.1037/dev0001752>
- Geva, S. & Fernyhough, C. (2019). A Penny for your Thoughts: Children's Inner Speech and its Neuro-Development. *Frontiers in Psychology* 10, Article 1708. <https://doi.org/10.3389/fpsyg.2019.01708>
- Goddu, M. K. & Gopnik, A. (2020). Learning what to change: Young children use "difference-making" to identify causally relevant variables. *Developmental Psychology*, 56(2), 275-284. <https://doi.org/10.1037/dev0000872>
- Hurlburt, R. and Akther, S. (2008). Unsymbolized Thinking. *Consciousness and Cognition*, 17(4), 1364-1374. <https://doi.org/10.1016/j.concog.2008.03.021>
- Jorba, M. & Vicente, A. (2014). Cognitive Phenomenology, Access to Contents, and Inner Speech. *Journal of Consciousness Studies*, 21(9-10), 74-99.
- Kimura, K. & Gopnik, A. (2019). Rational Higher-Order Belief Revision in Young Children. *Child development*, 90(1), 91-97. <https://doi.org/10.1111/cdev.13143>
- Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2), 622-646. <https://doi.org/10.1111/j.1467-8624.2007.01018.x>
- Lagenhoff, A. F., Srinivasan, M. & Engelmann, J. M. (2024). Disagreement reduces overconfidence and prompts exploration in young children. *Child development*, 95(5), 1616-1627. <https://doi.org/10.1111/cdev.14098>
- Lurz, R. W. (2007). In Defense of Wordless Thoughts About Thoughts. *Mind and Language*, 22(3), 270-296. <https://doi.org/10.1111/j.1468-0017.2007.00309.x>
- Martínez-Manrique, F. & Vicente, A. (2010). What the...! The role of inner speech in conscious thought. *Journal of Consciousness Studies*, 17(9-10), 141-67.
- Martínez-Manrique, F. & Vicente, A. (2015). The activity view of inner speech. *Frontiers in Psychology*, 6(232), 1-13. <https://doi.org/10.3389/fpsyg.2015.00232>
- Munroe, W. (2021). Reasoning, rationality, and representation. *Synthese*, 198, 8323-8345. <https://doi.org/10.1007/s11229-020-02575-6>
- Nedergaard, J., & Lupyan, G. (2023). Not Everyone Has an Inner Voice: Behavioral Consequences of Anendophasia. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45. <https://doi.org/10.1177/09567976241243004>
- Nedergaard, J., Wallentin M. & Lupyan, G. (2023). Verbal interference paradigms: A systematic review investigating the role of language in cognition. *Psychonomic bulletin & review*, 30(2), 464-488. <https://doi.org/10.3758/s13423-022-02144-7>
- Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science: A Journal of the American Psychological Society*, 18(7), 574-579. <https://doi.org/10.1111/j.1467-9280.2007.01942.x>

- Petrolini, V., Jorba, M. & Vicente, A. (2020). The Role of Inner Speech in Executive Functioning Tasks: Schizophrenia with Auditory Verbal Hallucinations and Autistic Spectrum Conditions as Case Studies, *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2020.572035>
- Vicente, A. & Jorba, M. (2019). The Linguistic Determination of Conscious Thought Content, *Noûs*, (3), 737-759. <https://doi.org/10.1111/nous.12239>
- Vygotsky, L. (2012/1934). *Thought and language*. MIT Press.
- Walker C. M., Bridgers S. & Gopnik A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. *Cognition* (156), 30-40. <https://doi.org/10.1016/j.cognition.2016.07.008>
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)

DAPHNE BERNUÉS MOSS is conducting her Ph.D. at the Humanities Department, Pompeu Fabra University, Barcelona. She focuses on the topics of inner speech and agency.

ADDRESS: Departament d'Humanitats, Universitat Pompeu Fabra, Barcelona. C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: daphnececilia.bernues@upf.edu ORCID: 0009-0006-7835-832X

MARTA JORBA GRAU is Serra Húnter Associate Professor at the Humanities Department, Pompeu Fabra University, Barcelona. She specializes in philosophy of mind and psychology and feminist theory.

ADDRESS: Departament d'Humanitats, Universitat Pompeu Fabra, Barcelona. C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: marta.jorba@upf.edu ORCID: 0000-0001-6484-2301



CAN QUASI-CYCLICAL PREFERENCES BE RATIONAL?

(¿Pueden ser racionales preferencias cuasi-cíclicas?)

Daniel Dohrn*

Università degli Studi di Milano

<https://orcid.org/0000-0002-0148-307>

Keywords

Preference
Decision
Rational
Rational choice
Action

ABSTRACT: One important dimension in the work of José Luis Bermúdez concerns broadening our ideal of deliberational rationality beyond the narrow confines of standard rational decision theory. In this vein, he (2020, 2022) presents two claims: (i) there are ‘quasi-cyclical preferences’. (ii) Quasi-cyclical preferences can be rational. These claims contradict received wisdom in decision theory. I shall critically discuss both claims. In particular, I shall question whether there are quasi-cyclical preferences, and whether they can be rational. As for the latter issue, I shall consider three paradigmatic applications: time inconsistency, game theory, and discursive deadlock.

Palabras clave

Preferencias
Decisión
Racional
Elección racional
Acción

RESUMEN: Una dimensión importante del trabajo de José Luis Bermúdez se refiere a la ampliación de nuestro ideal de deliberación racional más allá del estrecho marco de la teoría de la decisión racional estándar. En esta línea, Bermúdez (2020, 2022) presenta dos tesis: (i) existen «preferencias cuasi-cíclicas»; (ii) las preferencias cuasi-cíclicas pueden ser racionales. Estas tesis contradicen ideas tradicionalmente aceptadas en la teoría de la decisión. Discutiré críticamente ambas tesis. En particular, cuestionaré si existen preferencias cuasi-cíclicas y si pueden ser racionales. Con respecto a esta última cuestión, consideraré dos aplicaciones paradigmáticas: inconsistencia temporal, teoría de juegos y bloqueos discursivos.

Gako-hitzak

Lehentasuna
Erabakia
Arrazionala
Hautu arrazionala
Ekintza

LABURPENA: José Luis Bermúdezen laneko alderdi garrantzitsu batek erabaki arrazionalaren teoria estandarren muga estuetatik harago zabaltzen du deliberazio-arrazionaltasunaren gure eredua. Ildo horretan, bi baieztapen egiten ditu (2021, 2022): (i) «lehtasun kuasiziklikoak» daude; (ii) lehtasun kuasiziklikoak arrazionalak izan daitezke. Baieztapen horiek erabakiaren teoriaren jakinduria konbentzionalarekin kontraesanean daude. Kritiki aztertuko ditut bi baieztapenak. Batez ere, zalantzan jarriko dut ea badauden lehtasun kuasiziklikoak eta ea arrazionalak izan daitezkeen. Azken auzi horri dagokionez, hiru aplikazio paradigmatico aztertuko ditut: denbora-inkonsistentzia, joko-teoria eta diskurtso-blokeoak.

* **Correspondence to:** Daniel Dohrn. Dipartimento di Filosofia, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy – daniel.dohrn@unimi.it – <https://orcid.org/0000-0002-0148-307>

How to cite: Dohrn, Daniel (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>).

Received: 27 January, 2024; Final version: 06 November, 2024.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. What are quasi-cyclical preferences?

My aim is to critically discuss José Luis Bermúdez's proposal that there are quasi-cyclical preferences, and that they play a role in rational deliberation. Bermúdez introduces quasi-cyclical preferences by distinguishing them from *cyclical* or intransitive preferences. The latter are standardly taken to be irrational (2020, p. 79):

There are good reasons to think that it is irrational to have preferences that are cyclical. A decision-maker has cyclical preferences when, for example, she simultaneously prefers o_1 to o_2 , o_2 to o_3 , and o_3 to o_1 . A decision-maker with cyclical preferences *will never be able to decide to do what she prefers most (assuming that transitivity holds). For each of o_1 , o_2 , and o_3 there will always be something she prefers to it.* (2022, p. 3, m.e., notation adapted)

Cyclical preferences are defined as follows: an agent has cyclical preferences if, among her options, she prefers o_1 to o_2 , o_2 to o_3 , and o_3 to o_1 . Such preferences cannot properly guide decision-making. The following connection between preferences and rational decisions seems plausible: among any options $o_1, o_2, \dots, o_n$, it is rational to choose o_i only if there is no option o_j that one prefers over o_i . Yet cyclical preferences leave the agent in a quandary: whatever option she chooses, there will be one she prefers, i.e. would rather choose. Thus, there is no option that could be rationally chosen.

Bermúdez introduces *quasi-cyclical* preferences in three main steps (2021, ch. 6). In a first step, he claims that preferences do not range over outcomes but *framed* outcomes:

I suggest that the objects of preference are framed outcomes. There is no such thing as making choices over a purely extensional opportunity set, independent of any way of describing or framing the things in it. (2022, p. 3)

Preferences concern outcomes of one's actions given certain states of the world. Usually, preferences are assumed to concern outcomes as they are independently of how we represent them. In contrast, Bermúdez claims that preferences concern outcomes as represented or framed in a certain way.

In order to support this claim, Bermúdez points to framing effects that are well-established in the psychological research on heuristics and biases (2020, ch. 2). Against the widespread view that framing effects are irrational, Bermúdez emphasizes that they can be useful for decision-making. Given our limited purview of an indefinite number of options and their potential consequences, framing is a way of focusing attention on certain outcomes (2020, pp. 120-123). These arguments do not yet show that preferences themselves come framed, though.

Bermúdez takes a second step (2020, p. 81; 2022, p. 3). Comparable to opaque beliefs, framing the same outcome o_1 under two frames $F_A(o_1)$ and $F_B(o_1)$ opens up the possibility of preferring o_1 to a different outcome o_2 under one frame and preferring o_2 to o_1 under a different frame. For comparison, 'Hesperus' and 'Phosphorus' refer to the same celestial body, but we may refer to that body under the diverse descriptions 'Hesperus is the bright celestial body visible in the evening' and 'Phosphorus is the bright celestial body visible in the morning' (2020, p. 12). This opens up the possibility of rationally believing Hesperus to be distinct from Phosphorus. Quasi-cyclical preferences function analogously to such a case of *opaque* belief.¹ However, Bermúdez sees one important difference to the case of referentially opaque beliefs. Quasi-cyclical preferences are *referentially transparent*. $F_A(o_1)$ and $F_B(o_1)$ are known to be frames for the same outcome (2020, p. 98).

Bermúdez develops his case for quasi-cyclical preferences by careful analyses of classical fictional characters like Agamemnon in Aulis (2021, ch. 6) and Macbeth (2021, p. 108-112): "Macbeth prefers Fulfilling his Double Duty to Duncan [o_1] to Murdering the King [o_2]... Macbeth prefers Bravely Taking the Throne [o_2] to Backing Away from his Resolution to Make the

¹ A belief is opaque if the relationships of co-referentiality, e.g. between 'Hesperus' and 'Phosphorus', are not known to the believer.

Prophecy come True [θ_1].” (2022, p. 5) The example gives us some idea of what frames are like. The frames in the example include a descriptive component (keeping one’s allegiance to Duncan, killing Duncan, becoming king), but also something akin to a thick normative concept like fulfilling duty, murder, braveness. As we shall see, frames may also involve emotional components. On the one hand, it seems plausible to think of frames as holistic, potentially including anything that informs our attitude towards a certain outcome. Yet on the other hand, there are certain limits. For instance, one should not simply inscribe preferences themselves into the frames if an analysis of preferences as framed is to be informative.

In a third step, Bermúdez claims that quasi-cyclical preferences can be perfectly *rational*: “Such patterns of quasi-cyclical preferences can be *correct and appropriate* from the normative perspective of how one ought to reason.” (2022, p. 1, m.e.) Quasi-cyclical preferences can be adequate responses to diverging reasons. They are instrumental in dissolving time-inconsistencies, game-theoretical dilemmata, and discursive deadlock (see my section 3.1., 3.2.1., 3.2.2., 3.2.3., respectively). Bermúdez imposes a minimum requirement of rationality on framing outcomes: “...it seems rationally permissible to view situations and outcomes under multiple frames, provided that those frames are complementary and consistent (i.e. they do not contradict each other) and that none of them are based on false beliefs or comparably dubious sources.” (2020, p. 135) The requirement is that one ought not to accept frames which contradict each other or are based on false beliefs. In order to be rational, quasi-cyclical preferences must at least satisfy this minimum requirement. However, as Bermúdez’s examples of Macbeth etc. show, this minimum requirement leaves us with a wide variety of combinations that are rationally permissible.²

In sum, Bermúdez takes three steps away from received wisdom. Preferences come framed. There are quasi-cyclical preferences. Quasi-cyclical preferences can be perfectly rational.

2. Are there quasi-cyclical preferences?

2.1. ARE PREFERENCES FRAME-RELATIVE?

Bermúdez opposes orthodoxy about preferences as it has been developed from pioneering works like (Arrow 1951, ch. 2; Savage 1954, pp. 13-16 as discussed in Bermúdez 2020, pp. 118-121) onwards. The orthodox view is defended against Bermúdez e.g. by Guala (2022; Lau 2022; Weisel and Zultan 2022). Bermúdez and his opponents agree that there may be quasi-cyclical *wishes* or *desires*. For instance, Macbeth wishes to be king, even if this means murdering the king (Guala 2022, pp. 27). At the same time, Macbeth may wish not to murder the king, even if this means not to be king for him (“Wake Duncan with thy knocking! I would thou couldst!”). But orthodoxy denies that *preferences* are frame-relative, and therefore that there are quasi-cyclical preferences. The difference between Bermúdez and orthodoxy can be summarized thus:

Orthodoxy: there are frame-relative emotional responses, reasons, desires, but preferences are not frame-relative.

Bermúdez: there are frame-relative emotional responses, reasons, desires, *and preferences*.

I shall consider one exemplary argument of Bermúdez’s against his orthodox opponents:

We both accept that there are frame-relative emotional responses, reasons, desires, and so forth. I argue, though, that it will not always be possible to turn those frame-relative reasons into frame-neutral all-things-considered preferences. This will happen when the force and appeal of the rea-

² Bermúdez assumes that standards for factual truth are highly context-relative. For instance, prophecies count as fact-stating for Agamemnon assuming his whole community of reference accepts them, but evangelical Christians’ belief in the Rapture does not count as factual given the standards of contemporary society (2021, p. 233-234). As a consequence, the requirement seems rather to be that a frame should not be based on beliefs that *seem* false by a communal standard.

sons are tied to the frame in which they emerge, so that stepping back from the frame weakens their hold. (Bermúdez 2022, p. 57)³

The main argument Bermúdez here provides for making preferences frame-relative is that our reasons for certain preferences are inextricably frame-relative. There is no way for these reasons to influence the formation of preferences if not by informing frame-relative preferences. If we were to remove the frame and consider an outcome directly, these reasons would lose their grip on the formation of preferences.

There are several uncertainties about this argument. The first is whether the point is a psychological or a normative one: is the claim just that some reasons would lose their force, or is it that we would fail to do justice to certain reasons which should be taken into account? There is a related uncertainty: what is the consequence to be drawn from the observation that certain reasons are frame-dependent? Is there anything lost if we fail to appreciate these reasons?

While I shall come back to these issues, I here only note the following: even if it is granted that some reasons have an impact only on preferences formed under frames and not on preferences with regard to unframed outcomes, it does not follow that preferences are framed. It only follows that, if preferences are formed in a frame-neutral way, certain reasons cannot play a role in forming them. To illustrate this highly principled point by the Macbeth example, one version that is consistent with the story is the following: assume for the sake of argument that Macbeth could not prefer *Not Killing Duncan* to *Killing Duncan* unless framed as *Keeping* vs. *Not Keeping his Double Duty*, but that preferences never come framed. It only follows that Macbeth cannot prefer *Not killing Duncan* to *Killing Duncan*, which is consistent with the story unless we make certain contestable assumptions about preferences.

Things would be different if reasons principally could only have an impact on preferences under frames: assuming that rational preferences must be sensitive to reasons, and that only preferences under frames are sensitive to reasons, it would be a requirement on the rationality of preferences that they are preferences under frames. It is not obvious that Bermúdez would be willing to endorse this stronger claim, though. On the one hand, Bermúdez says: “There is no such thing as making choices over a purely extensional opportunity set.” (2022, p. 3) On the other hand, he says “...it will not *always* be possible to turn those frame-relative reasons into frame-neutral all-things-considered preferences.” (2022, p. 57, m.e.) Not always, but presumably sometimes. Bermúdez does not unambiguously exclude rational frame-neutral all-things-considered preferences. In any case, the argument against the orthodox position does not support the stronger claim.

In sum, Bermúdez argues that some reasons can only have an effect on preferences if the latter are framed. The argument does not establish that preferences ever come framed. So far nothing excludes that preferences are always formed over an extensional opportunity set, although frames may be instrumental in pre-selecting the potentially indefinitely many options to be considered.

2.2. AGAINST CUASI-CYCLICAL PREFERENCES

Having articulated certain doubts about Bermúdez’s positive argument for *frame-relative* preferences, I shall now present two arguments against the claim that there are *quasi-cyclical* preferences. The first concerns how to define the notion *preference*. The notion has become a *terminus technicus* in the current debate on rational choice. It is more theoretical and less rooted in folk psychology than the concepts of desires, wishes, and will (see Hansson and Grüne-Yanoff 2022). I shall present a plausible first take on its theoretical role in terms of the following link to agency:

³ A clarification on all-things-considered preferences is in order. Bermúdez responds to Guala (2022) “...that a rational agent can have more than one maximally all-things-considered preference ordering, in cases where there are reasons, emotional responses, desires, and so on, that cannot be considered together.” (2022, p. 57) I think that *all-things-considered* preferences in Guala’s (2022) understanding would require precisely to *consider together* reasons etc. A more careful statement of Bermúdez view might be that one can have inconsistent sets of all-things-considered preferences, where all reasons etc. are considered or somehow represented but not all are *live* or *maximally salient* (on salience see Bermúdez 2020, p. 264; Pettigrew 2022, p. 36).

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 .

Sufficient Condition: one does not prefer o_2 over o_1 if one does o_1 rather than o_2 in a case conscious, deliberate, voluntary agency.

Quasi-cyclical preferences do not satisfy my two conditions. I shall raise two issues. The first issue regards the *Necessary Condition*: one cannot have a quasi-cyclical preference $F_A(o_1)$ over $F_B(o_2)$ and $F_C(o_2)$ over $F_D(o_1)$ and do o_1 rather than o_2 or o_2 rather than o_1 as a case of conscious, deliberate, voluntary action.

The second issue concerns the attribution of preferences. What guides us in telling whether someone prefers o_1 to o_2 or the other way round? How are we to identify preferences? For instance, how can we know that Macbeth prefers *Fulfilling his Duty* (o_1) to *Murdering the King* (o_2) although he acts against this preference, killing Duncan (o_2)?

One main criterion of ascribing preferences is overt action. As a first stab in this direction, I suggest the following heuristic criterion:

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.

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.? In sum, quasi-cyclical preferences raise an attribution problem. One may still uphold them as a theoretical posit, but the question becomes why to posit them in Bermúdez's examples like the Macbeth case rather than doing with the resources of orthodoxy.

In the case of Macbeth, Bermúdez mentions the following exemplary piece of evidence that may be used to motivate the attribution of preferences that conflict with manifest actions. Macbeth's feeling of remorse shows that he continues to prefer (i) *Fulfilling his Double Duty* (o_1) to *Murdering the King* (o_2) even if he acts otherwise (ii). Retrospective emotions like *regret* as experienced e.g. by Macbeth after murdering Duncan are evidence that quasi-cyclical preferences persist beyond executing their target actions: "One sign that decision-makers have retained quasi-cyclical preferences would be their experiencing regret or similar retrospective emotions." (Bermúdez 2020, p. 112, on the rationality of regret p. 265)

I agree that retrospective emotions like regret might be construed in terms of retaining quasi-cyclical preferences if one already grants such preferences. Macbeth regrets having murdered Duncan because under the frame *Fulfilling his Duty* he would have preferred and he still prefers not to murder Duncan. Yet I doubt that retrospective emotions provide independent evidence for quasi-cyclical preferences. On a more neutral ground, regret can as well be explained in terms of a change of preferences (Macbeth now prefers not to have murdered Duncan)⁴ or in terms of persisting conflicts of wishes or desires, which have to be resolved in order to form preferences (Guala 2022). In this alternative, Macbeth does not change his preferences, but his regret is explained by opposing wishes or desires.

⁴ I follow Pettigrew (2022, p. 36), who differentiates his view of Bermúdez's cases of quasi-cyclical preferences (e.g. Macbeth): if the relevant preferences are supposed to record judgements of *betterness*, the agent should be interpreted as being indifferent (no option being *absolutely* better). If they are supposed to explain behaviour, one's reasoning and one's behaviour manifest a shift from one set of non-quasi-cyclical preferences to another. Macbeth's killing Duncan displays an absolute preference for killing Duncan tout court over not killing him, and his subsequent regret displays a shift of the preference ordering towards not killing Duncan.

I shall address two more general concerns about my argument: First, it may sound overly behavioristic. Responding to Guala (2022), Lau (2022), Weisel and Zultan (2022), and Pettigrew (2022), Bermúdez (2022, Section R4.2) emphasizes that he thinks of preferences as psychological states. He explicitly rejects thinking of preferences first and foremost as revealed in overt actions (Bermúdez, 2022, pp. 56-57, see Samuelson 1938). Yet one can agree that preferences are psychological states and reject the revealed-preferences view while still maintaining a certain connection to agency as in my link to agency and my heuristic criterion. My suggestion imposes only weak constraints on the attribution of preferences. There may be plenty of non-revealed preferences. Of course, there is a danger that the debate becomes a merely verbal dispute, each party simply using ‘preference’ differently. To lead the dispute in a more constructive way, I prefer to ask what the most fruitful use of ‘preference’ is. As a consequence, my doubts become whether Bermúdez’s use is more fruitful than one that maintains the abovementioned links to agency.

A related worry is that my link to agency begs the question against Bermúdez, by requiring agents’ preferences to range over unframed outcomes rather than framed outcomes. Yet first, the supposed link to agency is compatible with the claim that preferences are framed. Overt actions may impose informative constraints on the attribution of preferences whether framed or not, even if they underdetermine frames. Moreover, I use the link to agency only to argue against the claim that there are quasi-cyclical preferences. Secondly, my dialectical aim here is not to refute Bermúdez’s conception by arguing from a common ground that Bermúdez would have to accept, but to elaborate a competing plausible view of preference. In doing so, I make assumptions that Bermúdez rejects. It then remains to be seen which conception is the most convincing one.

I think the action-guiding role described by my links to agency has to be filled, and the concept of a preference is the most convincing candidate for filling it. One may try to reconcile the action-guiding role with acknowledging quasi-cyclical preferences by assuming that the action-guiding role is limited to normal or ideal cases of preferences or to normal or ideal cases of agency, but that would require accepting that quasi-cyclical preferences are not the normal case.⁵

I have raised some doubts about the claim that preferences are frame-relative, and that they sometimes are quasi-cyclical. I shall now put aside these more principled doubts and consider whether quasi-cyclical preferences (if such there are) could be rational.⁶

3. *Are quasi-cyclical preferences rational?*

3.1. PRINCIPLED DISCUSSION

Again, I shall begin with critically assessing a positive argument of Bermúdez’s for the rationality of quasi-cyclical preferences and then discuss some general doubts of my own:

(H3) Framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation.

[...]

⁵ Bermúdez (p.c.) has responded that in the case of quasi-cyclical preferences our doing may fall short of conscious, deliberate, and voluntary agency. This move may assuage my doubts about the very existence of quasi-cyclical preferences, but it may be problematic with regard to their purported rationality. Conscious, deliberate, and voluntary agency is closely connected with categories like rationality and responsibility. If our action falls short of conscious, deliberate, and voluntary agency, it falls into a limbo between rational and responsible action and mere behaviour, as when someone acts under the influence of drugs or mental illness. It is not clear what happens in this case to the claim that quasi-cyclical preferences can be rational.

⁶ Doubts about their rationality have been articulated by Fisher (2022): whenever an agent realizes that the outcomes under different frames are the same, she already considers them in a frame-neutral way, incurring a rational commitment to form all-things considered preferences (see Bermúdez’s response 2022, p. 56).

But those quasi-cyclical preferences have emerged from a decision-maker seeking to satisfy the basic rationality requirement of doing justice to the complexity of the situation. They inherit the rationality of the process that generated them. (2022, p. 6)

On the one hand, Bermúdez here suggests a procedural criterion of rationality. It is sufficient for preferences to be rational to arise from a rational process. On the other hand, the rationality is not described in terms of criteria for due processing. Rather there is a direction of fit from our preferences to an independent situation. Sometimes quasi-cyclical preferences are the adequate response to the situation.

To assess this argument, I shall follow Bermúdez's use of examples from the literary classics. However, I shall consider an example that I find more apt than Bermúdez's examples of Macbeth and Agamemnon. The problem with the Macbeth example is that we have a contrast of very different frames, on the one hand a moral obligation and on the other hand a selfish ambition. The problem with the Agamemnon example is that an assumed obligation to appease the Gods by sacrificing one's daughter is alien to modern sensibilities.

I shall consider a different example, which a contemporary reader can better connect to. 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.⁷ 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)*.

The case illustrates how framing and reframing options can be a rational procedure for doing justice to a complex situation. The question is whether this is also a case in which the formation of quasi-cyclical preferences would be rational. Judging from Bermúdez argument, if it is *ever* rational to form quasi-cyclical preferences, it seems rational for Jean Valjean to form quasi-cyclical preferences in responding to the situation. He responds to an intractable moral quandary.

However, there are three caveats about this response. The first caveat concerns the hypothesis that preferences come framed. None of the examples, including the one from Victor Hugo, support the hypothesis that preferences come framed rather than the weaker hypothesis that framing has an influence on forming preferences, which then concern outcomes directly.

The second caveat concerns the hypothesis that Jean Valjean may form quasi-cyclical preferences. According to orthodoxy, it is not adequate to describe Valjean as having quasi-cyclical preferences. Rather he may be faced with conflicting reasons, demands, desires, wishes, and so on. So far there is no reason to ascribe specifically quasi-cyclical preferences to him. I shall put aside this caveat as I am granting for the sake of argument that there can be quasi-cyclical preferences.

The third caveat concerns more specifically the rationality of quasi-cyclical preferences. In reconstructing Bermúdez position, I tentatively distinguish between two distinct respects in which preferences can be labelled as 'rational'. The first may be called *input-rationality*, i.e. rationality of the processes that influence decision-making. 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.⁸

⁷ Hugo thinks of this situation as typical of the *condition humaine*. Valjean's nemesis Javert experiences a perfectly parallel situation in volume five.

⁸ I share the concerns of an anonymous reviewer as to whether it is useful to invoke quasi-cyclical preferences in explaining input-rationality: we seem to already have a number of terms to refer to the products of the preliminary phase of exploration – 'pro tanto' reasons, or 'partial

The second respect in which preferences can be rational may be called *output-rationality*. Output-rationality has to do with the action-guiding role of preferences as captured by the link to agency discussed in the last section. Quasi-cyclical preferences do not satisfy a corresponding rationality requirement that is acknowledged by Bermúdez himself. According to Bermúdez, cyclical preferences of o_1 over o_2 , o_2 over o_3 , and o_3 over o_1 are irrational because, for any decision o_1 , o_2 , and o_3 , there is an alternative that is preferred to it. Now the same goes for quasi-cyclical preferences. Consider again Jean Valjean's assumed preferences:

Serving the Community (not turning oneself in) trumps *Saving the Ruffian (turning oneself in)*. *Taking Responsibility for Past Offences (turning oneself in)* trumps *An Innocent Being Punished instead of Oneself (not turning oneself in)*. As a matter of fact, Jean Valjean decides to turn himself in. Yet by assumption, he prefers not to turn himself in to turning himself in under the frames *Serving the Community* and *Saving the Ruffian*. Analogously if he had taken the alternative decision not to turn himself in.

Generalizing, whenever one has quasi-cyclical preferences $F_A(o_1)$ over $F_B(o_2)$ and $F_C(o_2)$ over $F_D(o_1)$, 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. To use his own words, A decision-maker with quasi-cyclical "*preferences will never be able to decide to do what she prefers most (assuming that transitivity holds). For each [choice] there will always be something she prefers to it.*" (2022, p. 3)

Rational agency is blocked if there is no way of resolving quasi-cyclical preferences. Pettigrew (2022, p. 36) interprets Bermúdez as suggesting the following resolution (see e.g. Bermúdez 2020, p. 264): an agent with quasi-cyclical preferences may act as suggested by the most salient framing. For instance, Valjean might turn himself in because the framing *Saving the Ruffian* happens to be more salient than the framing *Serving the Community*. Pettigrew claims that this solution is better interpreted as settling for one coherent set of preferences (turning oneself in over not turning oneself in) rather than acting on quasi-cyclical preferences. While I remain neutral about Pettigrew's interpretation, I want to add two points: first, both the formation of quasi-cyclical preferences and the resolution of the resulting quandary for action are explained by one frame becoming momentarily more salient than the other. Such an explanation seems incomplete. For instance, it leaves open which factors lead to the situation in which Valjean just forms quasi-cyclical preferences, and which factors lead to the situation in which he already acts on the set of preferences made salient. Second, 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. At this point, one may resort to techniques of *salience management*: there are frames that should be considered. Hugo intimates that turning oneself in is the right decision. Similarly, Shakespeare intimates that killing Duncan is the wrong decision for Macbeth. However, the question becomes how to settle for the right framing, the one that should be raised to salience. If there is an answer to this question, and if the technique can be implemented, there is no rational role for quasi-cyclical preferences any more.

Taking stock, quasi-cyclical preferences may be rational in a minimal sense given by input-rationality. Yet they are inevitably irrational in a sense acknowledged by Bermúdez himself. They cannot be rational in guiding our decisions. I find it plausible (and it seems that Bermúdez agrees given his discussion of cyclical preferences) that full-blown rationality requires both input- and output-rationality. The point of preferences arguably *is to mediate the rational transition from reasons, desires, motives, wishes to action, from input rationality (doing justice to the complexity of the situation) to output rationality (doing justice to the requirements of rational agency)*. As a consequence, quasi-cyclical preferences could not be fully rational.

evaluations', for example. It is doubtful that there is any advantage in using the term preference at this stage, given that preferences in the technical sense are 'all things considered' evaluations. Yet I find it dialectically convenient to for the sake of argument grant input-rationality as a concession to Bermúdez in order to bring out my main point.

3.2. APPLICATIONS

Besides the principled discussion, Bermúdez presents three concrete examples where quasi-cyclical preferences may play a rational role. Since my argument from the last section was fully general, that rational role can at most lie on the input side. Still it is instructive to go through these three examples as they help us appreciate Bermúdez's position.

3.2.1. Time-inconsistencies

Bermúdez's first example is time-inconsistent preferences as exemplified by *hyperbolic discount functions* (2020, ch. 7; 2022, section 4):

Exponential discount functions: [...] the impact of a day's delay will be the same tomorrow as 25 years in the future. For that reason, exponential discounting is described as time consistent.

Hyperbolic discount functions are time-inconsistent, because the ratio of the discount function is not constant. The difference between having \$10 today and receiving \$11 tomorrow is much greater than the difference between having \$10 100 days into the future and having \$11 in 101 days. (2022, p. 6)

A hyperbolic discount function is characterized by a change in the discount function that tells you how much the value of a later reward decreases compared to that of a sooner reward. One standard explanation for hyperbolic discount functions is that *Smaller Sooner* rewards SS (10\$ today instead of 11\$ tomorrow) are *emotionally hot*, as contrasted to *Larger Later* rewards LL (11\$ tomorrow instead of 10\$ today). A reward of 10\$ today is emotionally hotter than a reward of 11\$ tomorrow, but a reward of 10\$ 100 days in the future leaves us as cool as a reward of 11\$ 101 days in the future.

Bermúdez suggests the following strategy of counteracting time-inconsistent preferences:

Agents can ensure that hot representations of SS are counter-balanced and kept in check by cooler representations that emphasize, for example, the long-term consequences of succumbing to temptation. Likewise, they can represent LL in ways that engage the hot system, thus steepening the LL discount function and preventing the SS discount function from crossing it. (2022, p. 7)

Bermúdez suggests a reframing strategy in terms of reversing the emotional 'temperature' by representing outcomes differently. One may frame LL as *Resisting Temptation* (LL). The rational result can be quasi-cyclical preferences, or so it seems: one prefers *Resisting Temptation* (LL) over *Succumbing to Temptation* (SS) and *Emotionally Hot* (SS) over *Emotionally Cool* (LL).

Coming to my assessment, I take the irrationality of hyperbolic discount functions for granted. I admit that reframing can be a perfectly rational strategy of overcoming time-inconsistencies. However, the same does not go for forming quasi-cyclical preferences.

First, one does not have to accept that preferences come framed to endorse the reframing strategy suggested by Bermúdez. The strategy may be implemented as follows: one starts from a preference for SS over LL because the former is emotionally hot. One then counters the effect of one's emotions by reframing SS as a case of succumbing to temptation and LL as a case of resisting temptation. The consequence of reframing is that one's preferences change so as to become time-consistent. One drops one's former preference of SS over LL and forms a new preference for LL over SS.

Second, by the same argument, even if one grants that there are framed preferences, there is no reason why the reframing strategy should not be performed completely without ever forming quasi-cyclical preferences. One reframes one's options so as to overcome the effect of one's original emotions in regulating their 'temperature'. If one succeeds, one will end with replacing preferences *Emotionally Hot* (SS) over *Emotionally Cool* (LL) by something akin to preferences *Resisting Temptation* (LL) over *Succumbing to Temptation* (SS).

Third, even if we accept the hypothesis that preferences come framed, quasi-cyclical preferences are not a rational reaction to the situation. The rational reaction would be to adopt a preference for *Resisting Temptation* (LL) over *Succumbing to Temptation* (SS) but to drop a preference for *Emotionally Hot* (SS) over *Emotionally Cool* (LL) that gives rise to time-inconsistencies.

Fourth, as we have seen, quasi-cyclical preferences cannot guide rational action. One cannot rationally choose LL over SS because one at the same time prefers SS over LL under the emotional frame. As a consequence, quasi-cyclical preferences do not serve to overcome time-inconsistency. Rather they leave one in a quandary. One retains the time-inconsistent preferences under the emotional frame. One admittedly may refrain from acting on them, but firstly, doing so is not a result of preference rationally guiding action given that one also prefers otherwise, and secondly, one may as well refrain from acting on the rational preferences. Since time-inconsistency is mainly a problem for one's agency, this shows that quasi-cyclical preferences are not a solution.

3.2.2. Game theory

Bermúdez's second application are cases from *game theory* like the following (2020, ch. 8-9, 2022, section 5): there is a car stuck in the snow. Each of the two inmates is faced with the option of digging or not digging. The pay-offs from each perspective are as follows:

Car stuck in the snow		
	Stay	Dig
Stay	0;0	4;1
Dig	1;4	2;2

A Nash equilibrium in game theory is a set of strategies such that each player's strategy is a best response (in terms of expected outcome for oneself) to the strategies of the others. There are two pure-strategy Nash-equilibria: 4;1 and 1;4. Given the other digs, you should stay. Given the other doesn't dig, you should dig. However, it has been observed that many people cooperate, i.e. choose the option that both dig (2;2). How can that be rational if the best response is different?

In order to explain how it can be rational to choose cooperation, Bermúdez suggests that reframing and resulting quasi-cyclical preferences may play a role. He distinguishes two frames:

'I'-frame:

In the "I"-frame, agents look only at their own payoffs, employing the type of best response reasoning that seeks a Nash equilibrium.

'We'-frame:

A team reasoner thinks about the payoff table from the perspective, not of an isolated individual, but instead from the perspective of a team member, or group member. (2022, p. 9)

Bermúdez distinguishes a self-centred *I*- and a fairness-oriented *We*-frame. Under the *I*-frame my preferences as an inmate are the following:

I

1. (1I) I stay, you dig (4;1) > (2I) I dig, You dig (2;2) >
2. (3I) I dig, you stay (1;4)
3. (4I) I stay, you stay (0;0)

Under the *We*-frame, my preferences are as follows:

We

1. (2WE) I dig, you dig
2. (1WE) I stay, you dig = (3WE) I dig, you stay
3. (4WE) I stay, you stay

In this picture, considerations of fairness lead me to reframe my options and to rationally develop quasi-cyclical preferences under an *I* and a *We*-frame.

Coming to my assessment, again I admit that reframing can be a perfectly rational strategy of sustaining cooperation in game situations. However, again the same does not go for forming quasi-cyclical preferences.

First, one does not have to accept that preferences come framed to endorse the reframing strategy suggested by Bermúdez. The strategy may be implemented as follows: one starts from preferences corresponding to the preferences under the *I*-frame, arising from self-centred considerations of maximizing one's outcome given the action of others. One counters the effect of self-centred considerations by considerations of fairness that lead to new preferences corresponding to those under the *We*-frame.

Second, by the same argument, even granting that there are quasi-cyclical preferences, there is no reason why the reframing strategy should not be performed completely without ever forming quasi-cyclical preferences. One reframes one's options so as to replace preferences under the *I*-frame by preferences under the *We*-frame.

Third, even if we accept that preferences come framed, nothing guarantees that quasi-cyclical preferences ever form the rational reaction to the situation. One may argue that considerations of fairness trump self-centred ones (or vice versa), or that both are incommensurable. In the former case, one should drop one's preferences under the *I*-frame (or vice versa). Even in the latter case, it does not follow that quasi-cyclical preferences are the right response. The rational reaction may be not to form any preferences, or retain the ones one has formed previously to adopting a new framing for want of a better option.

Fourth, as we have seen, quasi-cyclical preferences cannot guide rational action. One may end up with cooperating, but this is not a rational consequence of quasi-cyclical preferences. One may by parity of reasoning as well end up with not cooperating. Quasi-cyclical preferences do not serve to overcome lack of cooperation in game situations, but they leave one in a quandary.

3.2.3. Interpersonal conflict

Bermúdez's third application are situations of interpersonal conflict (2020, ch. 10, 2022, section 3):

[...] to tackle discursive deadlock it is not enough for a frame-sensitive reasoner simply to understand that a particular action or outcome can be framed in multiple ways. She needs to frame it herself in multiple ways simultaneously... It is at this point that rational framing effects can enter the picture. Different ways of framing, say, restrictions on gun ownership, are associated with different preferences. For that reason, someone who internalizes the competing frames that give rise to discursive deadlock will often end up with quasi-cyclical preferences. (2022, p. 13)

According to Bermúdez, reasons are often inseparably baked into frames. One cannot do justice to these reasons without to a certain extent adopting the frames, i.e. accepting the presuppositions that come with these frames. One cannot make the frames one's own without rationally responding to the reasons baked into them. The rational response to these reasons sometimes consists in forming preferences according to them. The resulting framed preferences put

together with the framed preferences one had before may lead to forming quasi-cyclical preferences. Such quasi-cyclical preferences are rationally formed in responding to reasons within frames. Sometimes such a procedure of endorsing frames so as to form quasi-cyclical preferences may be the rational way of dealing with situations of discursive deadlock and interpersonal conflict.

One of Bermúdez's examples is gun regulation (2020, p. 219; 2022, p. 11). A pro-gun activist prefers *Freedom (no regulation)* to *Interference (regulation)*, whereas an anti-gun activist prefers *Safety (regulation)* to *Increased Violence (no regulation)*. I shall try to illustrate Bermúdez's argument about the connection of reasons, frames, and preferences by developing the example: In order to understand how the pro-gun activist comes to have a preference *Freedom (no regulation)*, the anti-gun activist may have to accept the libertarian presupposition that no regulation means freedom from government interference, and that such freedom is valuable. Of course, one may argue for or against this presupposition, but such arguments have to stop somewhere. At some point, one may simply have to take the libertarian view for granted or the reject it altogether. The latter option may lead one to end in deadlock. The reason for opposing regulation is entangled with the libertarian presupposition. One cannot adopt the presupposition without thereby having a reason to oppose regulation, and that particular reason to oppose regulation depends on sharing the presupposition.

To overcome discursive deadlock, it may be necessary for the opponents to appreciate each other's framing. Sometimes doing so may require one to a certain extent to adopt the other's frames and include the resulting preferences into one's own system of preferences. Thus, the opponents may rationally end up with quasi-cyclical preferences, preferring both *Freedom (no regulation)* to *Interference (regulation)* and *Safety (regulation)* to *Increased Violence (no regulation)*.

Coming to my assessment, there are analogies to the examples considered in the last section, but there are also significant differences. I shall start with the differences: With regard to the preceding examples of time-inconsistency (3.2.1.) and game-theoretical dilemmata (3.2.2.), so far I have found no real argument why the rational role of frames ever has to go together with the formation of quasi-cyclical preferences. In contrast, Bermúdez's discussion of discursive deadlocks provides such an argument: sometimes one cannot do justice to reasons baked into a frame without letting the frame have an impact on one's preferences. To appreciate an opponent's position, one may have to endorse her frame together with its impact on the formation of preferences.

However, there is a further important difference: while the preceding examples of time-inconsistency and suboptimal outcomes in games give us concrete puzzles of rationality, discursive deadlocks are no puzzles of rationality. Time-inconsistency is assumed to be irrational. In games, individual rationality systematically leads to outcomes that are suboptimal by one's own lights. There is at least an appearance of irrationality in systematically and foreseeably ending with suboptimal outcomes (provoking Lewis's 'Why Ain'Cha Rich?', Lewis 1981). In contrast, discursive deadlocks are simply cases where different people disagree. This may lead to conflicts which one would prefer to avoid, but there is no systematic failure of rational behaviour in achieving the outcome that is best by one's own lights. There is no principled *rational* pressure to dissolve discursive deadlocks. It may sometimes be better to dissolve them, but then again it may not. Sometimes the only thing to do may be to agree that one disagrees.

Having outlined two significant differences to the preceding examples, I shall now outline commonalities. Two of the four points that I made for each of the preceding examples also apply to interpersonal conflict:

First, one does not have to accept that preferences come framed to endorse the reframing strategy suggested by Bermúdez. Switching among different frames is sometimes useful in solving discursive deadlock, but it may be implemented without forming framed preferences. One strategy that may lead to better mutual understanding is for the pro-gun and the anti-gun activist to put themselves into their opponents' shoes and to *simulate* the latter's frames without making them their own. Even if they make them their own to some extent, these frames may have an impact only on motives, desires, and so on, but not on one's preferences.

Admittedly, if Bermúdez is right about the connection between frames, reasons, and preferences, such options may only lead to a limited understanding of each other's positions. One cannot do full justice to the other's reasons, but one may

still do justice to them to a certain extent, avoiding the contagious effects of adopting an alien frame on one's own preferences. Yet it is not a matter of course that Bermúdez is right about the connection. If orthodoxy is right, preferences do not come framed. Moreover, it is not a matter of course that one cannot do full justice to reasons in the sense of appreciating their value and force without permitting them an influence on forming one's preferences. The opponents in the gun debate may completely understand the mindset of their opponents without feeling in any way moved to adapt each other's preferences. For instance, the anti-gun activist may gain understanding of her opponent simply by reflecting on the libertarian presupposition that freedom comes with lack of government interference, confronting it with a liberal (e.g. Rawlsian) conception of freedom, and so on.

Second,⁹ even if we accept the hypothesis that preferences come framed, and even if we accept that one cannot completely do justice to another person's reasons without developing quasi-cyclical preferences, I see no reason why adopting quasi-cyclical preferences would help dissolve interpersonal conflicts. Rather one would end up with adding an *intrapersonal* to an *interpersonal* conflict. Take the example of the pro-gun and the anti-gun activist. If only one of them ends up adopting the other's frame and resulting quasi-cyclical preferences, she could not rationally opt any more for or against safety regulations, there being always an option she prefers to the one she chooses. If she were to try to act according to her preferences, she would end in a quandary, giving her opponent the upper hand and betraying her own original political convictions. If both of them end up in this way, again there is no rational solution to the deadlock. They end up deadlocked within themselves. Of course, both may end with choosing the same option, for instance because the same frame happens to prevail in both cases, but only at the price of one of them failing to act according to her convictions. Alternatively, they may continue to disagree, remaining in discursive deadlock.

4. Summary

Taking stock, Bermúdez rightly highlights the importance of framing effects and of framing and reframing strategies. Framing and reframing surely is psychologically effective and can contribute to rational deliberation. Yet it does not follow that quasi-cyclical preferences exist, or that there is a rational role for quasi-cyclical preferences to play.

Acknowledgements

This research was funded by the Department of Philosophy "Piero Martinetti" of the University of Milan under the Project "Departments of Excellence 2023-2027" awarded by the Ministry of University and Research (MUR). I am grateful to José Luis Bermúdez and two anonymous referees for helpful comments.

REFERENCES

- Arrow, K. (1951). *Social choice and individual values*. John Wiley & Sons.
- Bermúdez, J.L. (2020). *Frame it again: new tools for rational decision-making*. Cambridge University Press.
- Bermúdez, J.L. (2022). Rational framing effects: a multidisciplinary case, Response to Commentators. *Behavioral and Brain Sciences* 45, 1-16, 51-59.

⁹ This point corresponds to the fourth point in the previous sections, respectively.

- Fisher, S.A. (2022). Defining preferences over framed outcomes does not secure agents' rationality. *Behavioral and Brain Sciences* 45, 24-25.
- Guala F. (2022). Consistent preferences, conflicting reasons, and rational evaluations. Rational framing effects: A multidisciplinary case. *Behavioral and Brain Sciences* 45, 27-28.
- Hansson, S.O., Grüne-Yanoff, T. (2022). Preferences. In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2022/entries/preferences/>.
- Lau, J. Y. F. (2022). Ceteris paribus preferences, rational framing effects, and the extensionality principle. *Behavioral and Brain Sciences* 45, 30-31.
- Lewis, D. (1981). Why ain'cha rich. *Noûs* 15, 377-380.
- Pettigrew, R. (2022). Competing reasons, incomplete preferences, and framing effects. *Behavioral and Brain Sciences* 45, 35-36.
- Samuelson, P. A. (1938). A note on the pure theory of consumers' behaviour. *Economica* 5, 61-71.
- Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons.
- Weisel, O., Zultan, R. (2022). Frames, trade-offs, and perspectives. *Behavioral and Brain Sciences* 45, 47-48.

DANIEL DOHRN is an assistant professor at the philosophy department of the state university of Milan, Italy. He specializes in analytic epistemology, philosophy of language, and metaphysics. He also has interests in decision theory and in the history of modern philosophy.

ADDRESS: Dipartimento di Filosofia, Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy, daniel.dohrn@unimi.it – ORCID: 0000-0002-0148-307



DISCURSIVE DEADLOCK AND THE SOCIAL (MEDIA) LIFE OF FRAMES

(Bloqueo discursivo y la vida social (mediática) de los marcos)

Sarah A. Fisher*

Cardiff University

<https://orcid.org/0000-0003-1115-6134>

Keywords

Framing
Frame
Framing effect
Judgement and decision-making
Discursive deadlock
Social media

ABSTRACT: Disputes over complex social and political issues frequently reveal themselves in the use of contrasting frames. A policy on restricting gun ownership may be characterised by one side as “improving public safety” and by the other side as “curtailing individual rights”; one group may frame the construction of new homes on the greenbelt as “increasing the housing supply” while their opponents characterise it as “damaging the environment”; and so on. In recent work, J. L. Bermúdez argues that these kinds of alternative frames put agents in touch with distinct —and potentially conflicting— reasons for judgement and action. He diagnoses many cases of discursive deadlock as clashes of frames and he prescribes to agents a course of epistemic due diligence, whereby they adopt multiple alternative frames. In this paper, I argue that the analysis, while valuable, is importantly incomplete. Frames not only elucidate reasons but signal identities and affiliations —and that calls instead for structural changes to discursive environments.

Palabras clave

Marco
Enmarcar
Efecto marco
Juicio y toma de decisiones
Bloqueo discursivo
Redes sociales

RESUMEN: Las disputas sobre cuestiones políticas y sociales complejas suelen manifestarse en el uso de marcos contrapuestos. Una política de limitación de la posesión de armas puede ser caracterizada por una parte como una medida destinada a «mejorar la seguridad pública» y por otra como una «restricción de derechos individuales»; un grupo puede enmarcar la construcción de nuevas viviendas en zonas protegidas como una manera de «incrementar la oferta habitacional», mientras que sus opositores la describen como una acción que «daña el medio ambiente»; y así sucesivamente. En trabajos recientes, J. L. Bermúdez sostiene que este tipo de marcos alternativos pone a los agentes en contacto con razones distintas —y potencialmente conflictivas— para el juicio y la acción. Bermúdez interpreta numerosos casos de bloqueo discursivo como choques entre marcos y prescribe a los agentes una forma de diligencia epistémica consistente en adoptar múltiples marcos alternativos. En este artículo sostengo que, aunque valioso, dicho análisis es significativamente incompleto. Los marcos no solo esclarecen razones, sino que también señalan identidades y afiliaciones; y ello exige, más bien, transformaciones estructurales en los entornos discursivos.

Gako-hitzak

Markoa
Enmarkazioa
Enmarkazio-efektua
Judizioa eta erabakimena
Diskurtso-blokeoa
Sare sozialak

LABURPENA: Auzi sozial eta politiko konplexuen inguruko eztabaidetan, aurkako markoak erabili ohi dira. Ar-mak edukitzeko eskubidea mugatzera bideratutako politika bati buruz alderdietako batek enmarka dezake «segur-tasun publikoa hobetuko» duela, eta besteak, berriz, «eskubide indibidualak murriztuko» dituela; talde batek en-marka dezake «etxebizitza-eskaintza handituko» dela gerriko berdean etxebizitza berriak eraikiz gero, eta aurkari-ek, berriz, «ingurumenari kalte egingo» zaiola; eta abar. Duela gutxi argitaratutako lan batzuetan, J.L. Bermúdezek dio mota horretako marko alternatiboek judizioak egiteko eta jarduteko askotariko arrazoiak —eta agian kontraesankor-rak— eskaintzen dizkietela agenteei. Diskurtso-blokeoko kasu asko markoen arteko talka gisa diagnostikatzeko ditu, eta diligentsia epistemikoko prozesu bat gomendatzen die agenteei, alegia, era askotako marko alternatiboak beren-ganatzea. Lan honetan azalduko dudanez, analisi hori, baliotsua izan arren, nabarmenki osatugabea da. Markoek, arrazoiak argitu ez ezik, identitate eta kideztasunak adierazi ere egiten dituzte; beraz, aldaketa estrukturalak egin behar dira diskurtso-inguruneetan.

* **Correspondence to:** Sarah A. Fisher. Cardiff University College of Arts, Humanities and Social Sciences, John Percival Building, Colum Drive, Cardiff, UK, CF10 3EU. – fishers15@cardiff.ac.uk – <https://orcid.org/0000-0003-1115-6134>

How to cite: Fisher, Sarah A. (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>).

Received: 26 April, 2024; Final version: 28 November, 2025.

ISSN 0495-4548 – eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. Introduction

Why might someone describe the restriction of gun ownership as “improving public safety” or, alternatively, as “curtailing individual rights”? Is it because they value different things (or the same things to different extents)? Is it because they are focusing just on one facet of the issue, blinkered to considerations on the other side of the argument? Would collective decision-making be improved by discussants incorporating alternative frames into their thought and talk, forcing them to engage with their opponents’ perspectives? It is certainly tempting to think so. In this paper, however, I will suggest that things are not quite so simple —especially in the social media age.

To prefigure the argument that follows, I will show how frames not only elucidate facts, reasons, and values, but also signal a speaker’s identity and affiliation with others. Hence, incorporating alternative frames into their thought and talk is not merely a question of undertaking epistemic due diligence but of signalling solidarity with an out-group —an individually risky strategy that may also be collectively ineffectual.

I begin by introducing the concepts of *framing* and *frames* (Section 2). I then summarise Bermúdez’s proposal that frames elucidate potentially conflicting reasons; and explain how this grounds his diagnosis of, and prescription for, discursive deadlock (Section 3). I critique his analysis by showing how frames can also signal identity and affiliation (Section 4) and explain how this complicates both the diagnosis of discursive deadlock and the prescription for treating it (Section 5). Finally, I illustrate the acuteness of the problem in social media, due to features like connectivity, publicity, and decontextualization, before sketching out some possible solutions (Section 6). In so doing, I aim to open up new avenues for improving debate and decision-making in the digital age.

2. Framing and 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*.¹

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 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.²

¹ In order to remain neutral on philosophical debates about description and direct reference, I prefer to talk in terms of alternative *articulations*.

² For further discussion of conceptions of ‘frame’ in academic literatures, see (Bermúdez, 2020, pp. 11-14). Bermúdez himself deliberately eschews a precise definition, instead characterising the core idea as “a schema of interpretation that allows individuals to locate, perceive, identify, and label” (Bermúdez, 2020, p. 14). This bears similarities to Elisabeth Camp’s conception of frames as expressing *perspectives* — where perspectives are understood as “open-ended dispositions to interpret, and specifically to produce intuitive structures of thought about, or characterizations of, particular subjects” (Camp, 2019, pp. 18-19). Given my interest here in *discursive* deadlocks, I focus narrowly on linguistic devices for talking about some individual, entity, or state of affairs.

Taking our cue from the discussion towards the end of Bermúdez’s book, we will home in still further on the presentation of *complex social and political issues*, different facets of which are captured by distinct frames (Bermúdez, 2020, Chapters 10–11). To take a few more examples (my own, not Bermúdez’s), participating in strike action might be characterised as “protecting workers’ rights” or “disrupting people’s lives;” building new homes on the greenbelt might be characterised as “damaging the environment” or “increasing the housing supply;” and online content moderation might be characterised as “an attack on free speech” or “a harm reduction measure”.

Again, notice that none of these example pairs of frames involve logical equivalence: for instance, “an attack on free speech” and “a harm reduction measure” are not in general co-extensive. Nevertheless, within a specified context, the frames may be used to characterise the same thing (in this case, online content moderation).

Notice too that in each pair of frames considered, one presents the issue positively, while the other presents it negatively. Returning to an earlier example, the prospect of improving public safety lends intuitive appeal to a policy of restricting gun ownership, while the prospect of curtailing individual rights cuts against such a policy. Indeed, we might find that our views on the issue switch as we apply first one frame and then the other —and perhaps this frame-induced oscillation persists, even while we know that one and the same issue remains at stake. As we will see next, Bermúdez (2020, 2022) seeks to explain such frame-dependent preferences by appeal to the different *reasons* made available by each frame.

3. *Reasons and reasoning*

Bermúdez (2020, 2022) argues that frames can articulate different features or facets of a decision problem, thereby furnishing agents with distinct and potentially conflicting reasons. He asks us to consider, for example, the moral dilemma facing Agamemnon at Aulis:

Agamemnon is leading the Greek fleet against Troy to avenge the abduction of Helen by Paris. While the fleet is becalmed at Aulis, the prophet Calchas interprets a portent —two eagles swooping down to kill and eat a pregnant hare. As Calchas interprets the portent, it reflects the displeasure of the goddess Artemis at the prospect of innocents being killed at Troy. The lack of wind has the same source. The only solution, says Calchas, is for Agamemnon to sacrifice to the goddess his own daughter Iphigenia. (Bermúdez, 2020, p. 7)

As Agamemnon agonises over what to do, Bermúdez glosses him as framing and re-framing the bringing about of Iphigenia’s death, first as “murdering his daughter”; then as “following Artemis’s will”. The first frame furnishes Agamemnon with a reason not to bring about Iphigenia’s death, on account of murder being bad (and murder of one’s child especially so). However, the second frame furnishes him with a reason in favour of doing so, corresponding to the good associated with appeasing the powerful Goddess.³

The idea, then, is that thinking about a dilemma through one or other frame can focus an agent on reasons to grasp one or other horn. Similar analyses are given of Macbeth’s dilemma in Shakespeare’s classic play (Bermúdez, 2020, pp. 108–111), more prosaic everyday battles of self-control (Bermúdez, 2020, Chapter 7), and some classic game-theoretic coordination problems (Bermúdez, 2020, Chapters 8 and 9).

³ Bermúdez goes on to argue that it may be rational for Agamemnon to prefer following Artemis’s will to failing his ships and people, while also preferring failing his ships and people to murdering his daughter, even while knowing that following Artemis’s will in this case *is* murdering his daughter. For critical discussion of this further claim, see (Fisher, 2022a, 2022b). For the purposes of the current discussion, I am only concerned with the idea that the frames articulate distinct reasons for action.

Drawing lessons from the cases of *intrapersonal* decision-making, Bermúdez (2020, Chapter 10) turns to *interpersonal* deliberation—and specifically to ‘discursive deadlocks’. He defines discursive deadlocks as “evaluative and prescriptive conflicts between individuals and/or groups” (Bermúdez, 2020, p. 220) where “standard tools for dispute resolution and collective decision-making are ineffective in securing agreement” (Bermúdez, 2020, pp. 220-221). While he doesn’t specify what these ‘standard tools’ are, the basic idea seems to be that disagreement remains, even once all relevant facts are on the table (see also Bermúdez, 2022, p. 11).

Bermúdez is especially concerned about the nature of current debate over hot-button political issues, such as abortion, gun ownership, immigration, and climate change (Bermúdez, 2020, Chapter 10). He observes:

We are living in times where this type of deadlock is particularly widespread. It has become a platitude for contemporary political and social commentators to observe partisan deadlock and polarization on what are euphemistically called “values issues”. (Bermúdez, 2020, p. 221)

Let us consider in a little detail his more specific case of pledging allegiance to the American flag:

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)

Rather than giving us a neat pair of contrasting linguistic frames here, Bermúdez merely provides some suggestive associations and images corresponding to the opposing stances on pledging allegiance to the flag. Still, we can extrapolate from his discussion the terms in which the interlocutors might characterise the issue in debate. The political candidate might characterise the pledge as “reinforcing universal liberal values” (particularly in light of the last five words of the pledge being “liberty and justice for all,” as Bermúdez notes on p. 236). Their opponent might frame it as “reinforcing exclusionary nationalism or sectarianism” (indicated by the phrase “one nation, under God,” which also appears in the pledge).

While the good associated with reinforcing universal liberal values constitutes a reason to endorse and continue the practice of schoolchildren pledging allegiance to the flag, the bad associated with reinforcing exclusionary nationalism or sectarianism constitutes a reason not to. Hence, in discussing the issue through one or other frame, each agent is put in touch with a reason for or against the policy. Although they may agree on all the relevant facts (the wording of the pledge, the current state of US law and schools’ practices, and so on) they fail to converge on a policy solution. Indeed, the interlocutors may seem simply to talk past each other.

Bermúdez argues that the impasse need not be the result of a clash of (potentially incommensurate) values. For instance, in the case of the pledge, it might not be that one party endorses affirmation of universal liberal values while the other does not; nor that one party endorses affirmation of exclusionary nationalism or sectarianism while the other does not. Rather, Bermúdez suggests, the discursive deadlock may be better understood as a *clash of frames*. The adoption of one or other frame is what leads each party in the dispute to attend to distinct reasons.

This diagnosis of discursive deadlock as a clash of frames naturally invites a particular kind of prescription. Bermúdez argues that interlocutors ought to engage in “frame-sensitive” or “non-Archimedean” reasoning (Bermúdez, 2020,

Chapter 11). This involves agents first recognising that their opinion is shaped by their frame (a process Bermúdez calls “reflexive decentering” (Bermúdez, 2020, p. 243)) and then adopting their opponent’s frame (Bermúdez dubs this “imaginative simulation” (Bermúdez, 2020, p. 244))—but without discarding their original frame (thus they exhibit “perspectival flexibility” (*ibid.*)).⁴

In this way agents fulfil their “due diligence requirement,” according to which “a model frame-sensitive reasoner should make appropriate efforts to understand multiple ways of framing complex decision strategies” (Bermúdez, 2020, p. 271). And only then, with both frames in play, does interpersonal agreement become possible (for example, through convergence on the appropriate weight to be assigned to distinct frame-sensitive reasons). In a nutshell, the proposal is that participants in collective decision-making have epistemic obligations to work with multiple, potentially clashing, frames.

The approach is appealing. However, we might be sceptical that discursive deadlocks will, in fact, be broken, even once the disputants are fully au fait with others’ perspectives. Perhaps the process will still bottom out in a clash of values (for example, if views really do differ as to how valuable universal liberal values are, or how detrimental exclusionary nationalism or sectarianism is). Indeed, an interlocutor’s very choice to use one or other frame might reflect their already having considered the reasons at stake and come down on one side of the argument. Further due diligence would then seem fruitless.

I wish to pursue a related line of objection—one which challenges the idea that frames (always or only) put agents in touch with *reasons*. In the next section, I will explore another function of frames. Drawing on analyses in the social philosophy of language, I will argue that frames can be powerful signals of identity and affiliation. I will then show how this point complicates both Bermúdez’s diagnosis of discursive deadlock and his prescription for treating it.

4. Identity and affiliation

The insight I will be drawing on is that particular linguistic forms can conjure up certain individuals, social groups, or culturally salient characters. Consider, for example, a catchphrase like ‘I’ll be back,’ which calls forth Arnold Schwarzenegger’s Terminator, or a political slogan to make one’s country ‘great again,’ which conjures up Donald Trump and his supporters. This sort of effect is achieved by the speaker seeming to co-opt the words of another, speaking in their voice as it were.

To give a feel for how such uses of language construct and convey speakers’ identities and affiliations, I begin by introducing two extant analyses—of slurs and dogwhistles—as a run-up to the idea that framings of complex social and political issues can perform a similar function.⁵

4.1. SLURS

In recent years, linguists and philosophers of language have grappled with the problem of how exactly to characterise the derogatory nature of slurs, which appear to disparage people based on characteristics such as race, religion, or sexual

⁴ Here the language comes very close to that of Camp (2019). Accordingly, we might think of adopting an opponent’s frame as *inhabiting their perspective*, in Camp’s sense, with all of the concomitant attentional, emotional, and motivational implications that she describes.

⁵ For further philosophical discussion of how alternative linguistic forms carry such “social meaning”, see Nowak (2022; 2025) and Camp & Nowak (2025). For a book-length discussion of how these sorts of “resonances” might ground the very meanings of linguistic expressions themselves, see (Beaver & Stanley, 2023). I note too that, while my focus here remains on language, other expressive forms could, in principle, have similar effects.

orientation. There are many distinct approaches, appealing respectively to expressions' truth-conditional semantic contents (e.g. Hom, 2008), presuppositions (e.g. Schlenker, 2007), attitudes expressed (e.g. Richard, 2008; Jeshion, 2013), pragmatic implicatures (e.g. Bolinger, 2017), or taboos (e.g. Anderson & Lepore, 2013a, 2013b).

Our focus here will be on the account put forward by Nunberg (2018) and broadly endorsed by Pullum (2018). In 'The social life of slurs,' Nunberg argues that the pejorative import of slurs is explained by the fact that they are the terms used by bigots, whereas other members of society avoid using them. As he writes in relation to racist slurs: "racists don't use slurs because they're derogative; slurs are derogative because they're the words that racists use" (Nunberg, 2018, p. 244).

Nunberg denies that slurs express specifiable prejudiced or discriminatory content about their targets, either as a matter of semantic encoding or pragmatic communication; and nor does he believe they express any specific discriminatory attitude held by the speaker. Yet he maintains that slurs certainly do derogate their targets. According to Nunberg, this derogation is achieved purely through the association between the linguistic forms and those who use them: slurs are predominantly used by —and therefore strongly associated with— bigots.⁶

In elaborating the view, Nunberg uses the terminology of 'ventriloquism,' 'impersonation,' and 'material obtrusion' to describe how the use of a slur term seems to interpolate the voice of the bigot into the discourse. It is in this way that the bigot's discriminatory ideology enters the scene, somewhat indirectly and amorphously.

Nunberg further emphasises how the use of a slur term *affiliates* the speaker with the bigots it evokes, issuing a "declaration of solidarity with the speakers who own the word" (Nunberg, 2018, p. 286). In this way, a person's lexical choice ends up forging a nefarious alliance.⁷

In sum, Nunberg's analysis shows how a speaker's choice of one linguistic form over another (specifically, a slur term over a neutral term) signals a particular affiliation (in this case, with bigots).

4.2. DOGWHISTLES

Our second case study concerns 'dogwhistles,' i.e. uttered expressions interpreted in systematically different ways by distinct subsets of an audience. To take a few examples, Saul (2018, pp. 362-363) explains how the phrase "wonder-working power" (as used in President George W. Bush's 2003 State of the Union address) is likely to be heard by Christian fundamentalists as referring to the power of Christ, while most others would take it to be merely "an ordinary piece of fluffy political boilerplate" (Saul, 2018, p. 362). Stanley (2015) gives the example of "inner-city" —an apparently innocuous descriptive term that has become a racist dogwhistle in American political discourse. Khoo (2021) observes how "globalist" and "nationalist" have become white supremacist code words.

Discussing the last pair of terms, Khoo writes:

Over the course of the last century, these words have taken on racial and religious connotations, with "globalist" becoming associated with Jewish liberal conspiracy theories (many involving Hungarian-American banker George Soros), and "nationalist" associated with white Christian conservatives, in particular, the ideology of white nationalism: a loose coalition of commitments encompassing belief in white racial superiority, a fear of waning cultural and demographic dominance of white people in America, and the goal of dividing nations by race and establishing a white America [footnote omitted]. (Khoo, 2021, p. 147)

⁶ Even if one is sceptical about whether this is the *sole* way in which slurs derogate, Nunberg gives us a compelling account of *one* way in which they might do so. My thanks to a reviewer for encouraging me to be clearer about this.

⁷ Indeed, the affiliative feature of slurs is supposed to contribute importantly to their uniquely menacing force, since the user of a slur invokes the wider mob with whom they are identifying. As Nunberg puts it, "The word can turn a bigot from a hapless, inconsequential 'I' to an intimidating, menacing 'we'" (Nunberg, 2018, p. 286). See (Barnes, 2023) for related analysis of how collective authority is constituted online.

Importantly for our purposes, Khoo argues that the connotations of “globalist” and “nationalist” have become established purely via patterns of use among the relevant communities of speakers. For instance, he argues that the word “nationalist” is known to be used by members of the white supremacist community to identify themselves and other white supremacists (Khoo, 2021, p. 153). So, someone who endorses a “nationalist” position might thereby signal their affiliation with the white supremacist community.⁸

Two notes of caution are in order. First, Khoo doesn’t think coded speech *always* works through this metalinguistic mechanism; sometimes there is a more direct association between the meaning of an expression and what it dogwhistles, as when “inner-city” becomes a racist dogwhistle due to people’s beliefs about the racial demographics of inner cities (Khoo, 2017).

Second, dogwhistles can function somewhat similarly to slurs, with expressions like “globalist” and “nationalist” evoking bigoted individuals or groups and their vicious ideologies. However, they do so much more equivocally. It is still possible in many contexts to use terms like “globalist” and “nationalist” neutrally, as when someone condemns a “nationalist” position on account of its indifference to events in other parts of the world. Likewise, many competent and reasonable audience members may not hear the dogwhistle at all, even when it is intended or interpretable as such. It is far harder to imagine how a slur term could be used without derogation (at least for non-members of the targeted group—a point also made by Nunberg and others).

All of that said, Khoo’s analysis suggests that, in some contexts, speakers who endorse “nationalist” positions at least run the risk of appearing to identify with white supremacists. Once again, this is because particular linguistic expressions bring into play particular communities of speakers who use them in particular ways. Use of the expressions can then signal affiliation with those communities.

In discussing slurs and dogwhistles, we have been focusing mainly on ideologies outside of reasonable democratic discourse. However, we can see similar mechanisms at work *within* reasonable democratic discourse. This brings us back to alternative framings of complex social and political issues.

4.3. THE SIGNALLING FUNCTION OF FRAMES

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. As I will show, the prospects for breaking discursive deadlock in that way then seem rather slim, prompting us to pursue other solutions. Let’s now take the argument step by step.

First, drawing on the accounts of slurs and dogwhistles surveyed above, linguistic frames that articulate different facets of a complex issue might not *only* or *mainly* do that. For example, framing the restriction of gun use as “improving public safety” or “curtailing individual rights” does not only bring into play distinct features of the policy, and reasons to support or reject it. The first frame additionally evokes US Democrats (who typically endorse gun control in these terms) while the second frame evokes Republicans (who are more likely to characterise the issue that way). In many contexts, then, the choice to use one or other frame in public debate will signal allegiance to these political camps and their ideas (although the signal is defeasible, as when the frame is obviously being used in parody or counter-argument).

⁸ Henderson & McCready (2024) develop a somewhat similar account of dogwhistling. They draw on recent sociolinguistic research, which recognises that linguistic variation is not merely a function of individuals’ membership of social categories like gender, race, and class, but can be involved in constructing fine-grained personas (Eckert, 2012, 2024).

To take another example, consider discourse surrounding online content moderation. Someone who characterises it as “an attack on free speech” in the mid-2020s might well evoke right-wing public figures for whom the frame features as a central component of their rhetorical repertoire. In contrast, an opponent who endorses it as “a harm reduction measure” will tend to conjure up a more left-leaning character. By using the frames in discussion, speakers tend to signal their affiliation with different political groupings. Such speech effects are clearly separable from what the frames themselves articulate—and the reasons for judgement and action that they elucidate.⁹

Moving to the next step in the argument, I now show how the signalling function of frames complicates Bermúdez’s analysis of deliberative deadlock.¹⁰

5. *Discursive deadlock revisited*

Recall that Bermúdez suggests many apparent clashes of value may be better understood as clashes of frame. His idea is that interlocutors’ adoption of alternative frames leads to their attending to conflicting reasons, which in turn leads to their taking up opposing stances on contentious issues. Building on the discussion of the previous section, I will argue, first, that interlocutors may often have taken up opposing frames and stances because of their existing affiliations; and, second, that the adoption of alternative frames in such scenarios is unlikely to aid collective decision-making. I take each point in turn in the next two subsections.

5.1. DIAGNOSIS FOR DEADLOCK

In practice, both one’s stance on a contentious issue and one’s adoption of a particular frame may often be driven by one’s prior identification with others.¹¹ For example, one’s pre-existing affinity with the political right might lead one to approach a debate on online content moderation by adopting a frame they use—say, by railing against content moderation as “an attack on free speech”. Doing so has the benefit of at least appearing to make a substantive contribution while also, crucially, publicly reinforcing one’s social and political positioning. The idea here is that interlocutors are sometimes focusing at least as much on the social signals sent by their linguistic choices as on the informational content of the words themselves.¹²

This changes the diagnosis of discursive deadlock. The reason why we see intransigent exchanges of conflicting frames in interpersonal deliberation could be because people are outwardly reaffirming their distinct affiliations and alliances—not because they have failed to appreciate their opponents’ reasons. If this is correct, we should expect deadlock to persist in discourse, even among those who have engaged in epistemic due diligence in their private reflection.

⁹ Similar analyses can be given for other hot button political issues cited by Bermúdez, such as abortion, immigration, and climate change. Indeed, as political views on such issues become ‘stacked’ in the sense that members of society increasingly divide along the same lines, into the same political groupings, rather than these cutting across one another, an individual’s stance across a whole suite of issues may become ever more clearly signalled—see Mason (2018) for related discussion.

¹⁰ While my proposal, like Bermúdez’s, has affinities with Camp’s (2019), I believe there are at least two important points of difference. First, Camp seems to build into the definition of a frame that it has the function of expressing a perspective. In contrast, I define frames merely as linguistic alternatives, which may or may not express different perspectives. Second, on Camp’s view, the perspectives expressed by frames are *sets of dispositions*. In contrast, I am arguing that frames can *evoke individuals, groups, or culturally salient characters*, and thereby *signal affiliation*. See also Camp & Flores (2024)—including for discussion of the relationship between perspectives and personae. Thank you to a reviewer for pressing me on this point.

¹¹ For further related discussion on alliance-led formation of political views, see (Pinsof *et al.*, 2023).

¹² Malinowski (1923) refers to the role of speech in social bonding as its “phatic” function. The notion is picked up in recent analysis by McDonald (2021). See also (Keiser, 2023) for an excellent book-length treatment of how linguistic communication often departs from the communicative ideals around which philosophers have often centred their analyses.

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. Insofar as polarised public debate acts as a barrier to good collective decision-making about contentious issues, we might then expect the problem to persist in the face of further epistemic due diligence.

Of course, the empirical question remains: what proportion of the discursive deadlocks we witness are, in fact, of the clashing identities variety, rather than the clashing frames variety? While I do not have a definitive answer, I proceed on the assumption that the proportion is significant enough to require our theorising to accommodate it.

5.2. PRESCRIPTION FOR DEADLOCK

One might hope that even if Bermúdez's diagnosis of discursive deadlock is incomplete, his prescription is still the right one: speakers ought to incorporate their opponents' frames within their thought and talk. After all, based on the preceding analysis, this might not only allow them to appreciate each other's reasons but perhaps also to inhabit, in a richer sense, each other's social identities. For example, if a Democrat were to adopt Republican framing on an issue, perhaps that would enable them not only to reconsider arguments against policies like gun control, but also to experience *being* a Republican (including, perhaps, how others relate to them in that social position). This points to an even more comprehensive due diligence that is not only epistemic but psychosocial. Perhaps that would grant valuable empathetic insight into opponents' perspectives.

While this is an intriguing proposal, it comes with downsides. One is the potentially destabilising psychological effect of operating with multiple different self-identities (insofar as that is possible at all). Another is that introducing opponents' frames into one's speech is socially risky. Interpolating the voices of particular individuals or groups into a conversation might be wholly unwelcome and resisted by one's interlocutors. For instance, someone who appears to affiliate themselves with Republicans when in conversation with other Democrats, say by characterising the restriction of gun ownership as "curtailing individual rights," may incur disapprobation from their in-group. In this way, practising perspectival flexibility out loud is liable to send signals that harm one's social standing and relationships with others.

One possible response here would be to bemoan the epistemic timidity of those who eschew their opponents' frames merely to avoid creating social friction.¹³ Perhaps we should demand that rational agents adopt those frames, regardless of the risks.

While I have some sympathy with this response (and will later suggest ways to foster epistemic virtue more systematically within our discursive environments) I think the point can be overdone. After all, humans are social creatures whose reputations and relationships shape the course of our lives. While theorists can certainly try to understand and mitigate certain social pressures, simply willing them away is probably neither effective nor desirable. Better to acknowledge the value in social bonding, and recognise that linguistic communication is one means through which we achieve it, with speech therefore serving a mixture of epistemic and social functions.

The considerations above suggest that adopting opponents' frames in public discourse may not always be practical, even if it is desirable from an epistemic or ethical point of view. More importantly, however, doing so may not even be *effective*. Insofar as discursive deadlock is really a matter of upstream group affiliation, not frame-dependent reasoning, it is not clear why encouraging speakers to adopt their opponents' frames would change the underlying dynamics of the debate. Temporarily inhabiting an opponent's perspective could leave one's underlying social identity and affiliations untouched (or lead to some disconcerting flip-flopping between them). The result: debate proceeds in as polarised a manner as ever and results in either paralysed or uncompromising decision-making. In that case, Bermúdez might simply be giving us the wrong normative guidance when he encourages us to adopt our opponents' frames.

¹³ For related discussion of intellectual timidity, see Tanesini (2018).

In the final part of the discussion, I will illustrate the problem by turning the spotlight on social media, showing how it arises especially acutely there, before proposing an alternative remedy.

6. *Social media*

We have seen how linguistic frames might reveal divergent identities and alliances, not (only) conflicting reasons for judgement and action. I now show how this comes into sharp focus on social media platforms, our public discussion forums *du jour*. I argue that users face incentives to send unambiguous signals about their social positioning, potentially outweighing their incentives to exchange reasons. The discussion addresses three factors in turn —connectivity, publicity, and decontextualisation— that together render the adoption of alternative frames increasingly infeasible and ineffective.

6.1. CONNECTIVITY

Early social media platforms sought to replicate or reconstitute users' real life social networks online. In the late 1990s and early 2000s, sites like Classmates.com (in the US) and Friends Reunited (in the UK) enabled people to reconnect with long-lost friends and acquaintances. Facebook (originally set up in 2004) ultimately became the largest platform connecting users with family, friends, and acquaintances.¹⁴ For over a decade, the stated mission of Meta (which owns Facebook, Instagram, and Threads) has been to bring people together to build community (currently formulated as: "Build the future of human connection and the technology that makes it possible.")¹⁵ Thus, the focus has been very much on the *social* in social media. These platforms are spaces for people to maintain bonds with others. Users achieve this by participating in discourse, posting and reposting content, and replying and reacting to others' content.¹⁶

Social media is therefore quite unlike traditional media, such as newspapers, radio, and television, where content produced by professional journalists and broadcasters is transmitted to the general public. While traditional media outlets established themselves as forums for reporting and discussion of complex social and political issues, social media platforms were not designed to fulfil that function —at least not primarily or explicitly. That they have nonetheless ended up acquiring it, often displacing, subsuming, or fundamentally reshaping traditional outlets, presents society with various challenges. One of these, I suggest, is that users' speech often continues to prioritise social goals over epistemic ones.

Assuming that the goal of social bonding tends to be furthered by reinforcing shared identities and alliances, it is natural that we would aim to attract approval from in-groups —and that this would be reflected in our linguistic choices. For example, we can expect a left-leaning user to adopt forms of expression associated with those on the political left, while avoiding those associated with right-wing opponents. This is one way to help signal allegiance to, and bond with, other left-leaning contacts.¹⁷ In a space where bonding with allies is key, there is every reason to use in-group frames and little pressure to use those of out-groups. Other things being equal, then, we can expect frames associated with in-groups to be preferred over those associated with out-groups.

The proposal that in-group framing can support social bonding is complicated somewhat by a feature of social media known as 'context collapse' (Marwick & Boyd, 2010; Frost-Arnold, 2021; McDonald, 2025). This is the pooling

¹⁴ At the time of writing, Facebook is reckoned to have over three billion active monthly users. (<https://www.statista.com/topics/751/facebook/>, accessed 4th May 2025).

¹⁵ <https://www.meta.com/en-gb/about/company-info/> (accessed 4th May 2025).

¹⁶ While social media posts very often include non-linguistic content, such as images and videos, to keep things simple, the focus of the current discussion will remain on the linguistic components.

¹⁷ Frames are not the only – or probably even the best – way of doing so; others include profile information, followerships, and the contents of posts themselves.

within a single audience of contacts from different parts of one's social life, who potentially have quite discordant identities and affiliations. The idea is developed from the work of Goffman (1969). Goffman observes that people present themselves differently across contexts, establishing distinct 'faces' in interactions with family members, friends, work colleagues, strangers on the street, and so on.¹⁸ While Meyerowitz (1986) observes how communication technologies like the telephone, radio, and television already began to disturb the neat separation of social domains, the emergence of social media seems to have all but obliterated any remaining boundaries. On a platform like Facebook, for instance, all of the audiences mentioned above can be present simultaneously.

In collapsed contexts, agents are confronted with the dilemma of which 'face' to present. Speaking in the voice of any given individual, group, or culturally salient character risks alienating some members of one's social network, even while strengthening bonds with others. Thus, we might expect users to behave rather cautiously, in order to avoid offence; and that could take the form of deploying relatively neutral frames, or even refraining from posting on certain issues altogether.

Alongside such strategies of withdrawal, however, we also see strategies of entrenchment. These involve crystallising one identity in the face of competing possibilities (or selecting one 'face' over others). One implication of being able to connect with indefinitely many people all around the world has been to enable users to explore different communities and identities—potentially finding preferred versions of themselves in the process. As a result, an individual might react to context collapse by actively strengthening their identification with certain parts of their network at the expense of others (perhaps also 'culling' legacy contacts who are not well-aligned).¹⁹

It must be noted that social media platforms have recently begun to move away from a connectivity model towards an entertainment model. Rather than encouraging users to forge and maintain social connections in order to populate their feeds with content, algorithms increasingly serve up whatever—and whoever's—content is expected to be found compelling. TikTok is a clear frontrunner here, with its aim to "inspire creativity and bring joy" through the targeted dissemination of short video content.²⁰ Other platforms have sought to incorporate similar functionality. On this newer model, the emphasis is on keeping people amused and engaged rather than socially connected. It remains to be seen how this shift might affect the analysis of online framing provided here. In the next two subsections, I consider two other factors driving the signalling function of frames that continue to apply widely across different platforms.

6.2. PUBLICITY

Having considered why users might be inclined to use online speech for social bonding, I turn now to the additional pressure their audiences exert on them in the same direction. There is often no principled limit on who can view a post once it is published (and a post can often be tracked down even after its deletion, for instance through the posting and reposting of screenshots). One upshot of this is that large numbers of people might share or comment on a post—whether in positive or negative ways. This exposes ordinary users to reputational effects, including (in the negative case) public castigation.

On the one hand, it can be relatively easy to foresee the backlash to a post, especially when it comes from political opponents. For instance, expressing allegiance to a particular electoral candidate, party, or policy can be expected to meet with attack from those on the other side of the political divide. Drawing that kind of fire might be 'priced in' to a piece of online speech (or even part of what motivates it, as in cases of 'rage-baiting') and might have limited effects on the user, or their audience.²¹

¹⁸ It is perhaps no coincidence that Goffman was also a progenitor of research on *framing*, understood as the organisation of our social experiences (Goffman, 1974).

¹⁹ Hence the emergence of epistemic bubbles and echo chambers online (Nguyen, 2020).

²⁰ See <https://www.tiktok.com/about> (accessed 4th May 2025).

²¹ In practice, the vicious nature even of anticipated backlash can itself be grounds for serious distress and withdrawal from debate—see (Brown, 2018) and (Barnes, 2023) for discussion of how hate and abuse is exacerbated online.

On the other hand, there can be intense social pain associated with being publicly rebuked by in-group members, with whom one *does* wish to be identified and associated, and among whom one does wish to maintain esteem.

Billingham & Parr (2019) argue that even when in-group norm policing is well-motivated, it can easily end up becoming disproportionate and produce more harm than good. As a case in point, they discuss a user's tweet of: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!"²² Discussing her case, they explain:

She meant this as an ironic comment about the attitudes of middle-class white Americans toward AIDS, but many observers considered it racist. [Footnote omitted] [The user] faced a barrage of online criticism, much of it threatening and abusive, and was fired from her job — all while she was on a plane, and so unable to defend herself or explain her tweet. (Billingham & Parr, 2019, p. 372)

This ill-judged tweet was severely punished both through social sanctions (in the form of online public shaming) and material ones (due to the loss of employment).²³

Of course, the case described above is extreme in various respects. However, lower-key examples, where users face backlash for transgressing (or appearing to transgress) in-group norms, are commonplace in social media. Once users have witnessed or experienced enough such interactions, it is understandable that they would take care over how they frame things.

I suggest, then, that informal social pressure online can serve as an important check on users' linguistic frames. We can expect users to select their framings of contentious social and political issues not only to proactively reinforce their social bonds (as per the previous sub-section) but also more defensively, to avoid social sanctions. For example, a Democrat might wish to avoid framing the restriction of gun ownership in terms of "curtailing individual rights," while a Republican might wish to avoid framing it in terms of "improving public safety". Doing so at least risks raising suspicion among their fellow partisans — and could in some contexts contribute to public chastisement, with unpredictable implications.

6.3. DECONTEXTUALIZATION

In addition to their publicity, it is also notable that social media posts are typically short (with many platforms imposing character limits) and presented individually within algorithmically-determined feeds that largely obscure conversational context. This greatly increases the risk of miscommunication and misinterpretation.²⁴

It is often unclear exactly what an individual post is about or what it is doing, and audiences often lack knowledge about the relevant background discourse or events.²⁵ Of particular relevance to the current discussion, a user's identity, affiliation, and political views may not be immediately obvious from such a limited snippet of their speech. Disclaimers, clarifications, and corrections are easily lost within longer threads. Given such a scant basis for conveying information, users must rely heavily on how the message is crafted. There is a strong incentive for socially pertinent information to be as clear as possible, avoiding risky ambiguity. This includes sticking to frames that clearly signal in-group affiliation and avoid the risk of signaling out-group affiliation. Again, then, the importance of sending a clear signal through a short decontextualised post adds to social media platforms' encouragement of the use of in-group frames.

²² The case is drawn from Ronson (2015). See also (Aly & Simpson, 2018) for related discussion.

²³ See also Aly & Simpson (2018) for helpful related discussion.

²⁴ It is true that some platforms' algorithms present posts chronologically, while others seek to optimise for engagement. Either way, though, the speech bears little resemblance to turns in a back-and-forth conversation. See Goldberg (2021) for related discussion.

²⁵ Davies (2025) discusses how this can result in radically different interpretations of online content.

I have argued that, in social media environments, what a frame communicates about one's social positioning can loom large relative to any reason it might elucidate. Crucially, such environments inhibit the kind of perspectival flexibility in interpersonal communication that Bermúdez calls for. On the contrary, adopting opponents' frames in one's public posts risks causing social friction and possibly incurring social sanction. This makes the performance of epistemic due diligence especially difficult.

Moreover, it is unclear how adopting opponents' frames in such environments would help resolve discursive deadlock. To the extent that the frame is performing a social signalling function, doing so might simply send a confused signal about the user's identity and affiliations, without helping opposing groups come together in support of reasonable collective decisions.

Of course, the responsible epistemic agent might still make efforts to adopt conflicting frames in their *thought*, even if not in their *talk* (or *posts*). In this way, they might attain a more nuanced understanding of complex issues, had that been lacking beforehand. However, nuanced understanding won't necessarily solve the wider social problem of deadlock in public discourse. In the absence of sufficient incentive to break out of polarising communicative dynamics, we end up with precisely the kind of impasse that motivated Bermúdez's analysis in the first place —with communicative interactions that fail to appreciate the complexity of interlocutors and their ideas; and with visibly bitter contestation of political stances.

If Bermúdez-style due diligence is not the (complete) solution, what is? Insofar as the problem of discursive deadlock lies upstream in people's prior formation of alliances and affiliations, one answer would be to try to loosen the grip of these somehow. However, I will end by remarking very briefly on some actions we could take downstream, to restructure our discursive environments.

6.4. DUE DILIGENCE ONLINE

Building on the analysis of the previous section, solutions to discursive deadlock might be found in addressing the dysfunctions associated with connectivity, publicity, and decontextualisation.

Beginning with connectivity, part of the problem was that sites originally designed for establishing and reinforcing social connections have ended up numbering among our most important spaces for debate (both in the sense that more and more of our speech seems to take place there, and that it has a disproportionately large political influence). One possibility, then, would be to (re-)engineer online spaces with explicitly epistemic aims, making social bonding a less prominent goal. Perhaps that could support constructive collective reasoning. Examples of online spaces that take steps in this direction include Reddit's 'change my view' community²⁶ and vTaiwan's use of Pol.is to facilitate large-scale conversations and consensus building.²⁷

Turning to *publicity*, perhaps posts should not be so readily available for such wide consumption. Enhanced functionality for constraining the audience could make nuanced discussion less reputationally risky. When one incurs disapproval (even when it comes from in-group members) it is surely easier to rebuild bonds and regain face in a more private setting than one with indefinitely many onlookers. Just as our private offline social interactions can afford us relatively forgiving 'safe spaces' to work through ideas, it might be possible and desirable to recreate these dynamics online.

Moves already made in this direction include options for users to delete their previously published content, restrict the audience for posts, or switch off comments. Platforms could also experiment further with functionality for re-posting and 'quote-posting' (whereby users repost another's content overlaid with their own commentary), consider options for

²⁶ <https://www.reddit.com/r/changemyview/>, accessed 4th May 2025

²⁷ <https://info.vtaiwan.tw/>, accessed 4th May 2025

how rapidly content is permitted to spread across networks, or adopt more nuanced policies on anonymity or pseudonymity for users.

A different sort of approach would target *decontextualization*, to make the wider context of a post easier to access and engage with. The aim here would be to mitigate the risk of misinterpretation, such that users feel a little freer to pursue compromise without appearing to ally themselves with out-groups. This could involve changing the platform interfaces to make interactions more genuinely conversation-like.

The avenues sketched out here will no doubt be littered with obstacles, both of a principled and practical nature—and they are certainly not intended to be exhaustive solutions. The underlying point still stands, though: if we are to address discursive deadlock, it may be misguided simply to call for greater epistemic due diligence. Instead, we may need to attend more closely to the structures of our discursive environments, including social media platforms. For, as long as frames are used to signal opposing affiliations, discursive deadlocks seem likely to remain and inhibit good collective decision-making. To support constructive communicative behaviours, we should therefore design environments that facilitate, rather than impede, epistemic virtue.

7. Conclusion

I have suggested that the alternative linguistic frames used to characterise complex social and political issues do not only elucidate reasons for judgement and action, as Bermúdez argues, but can also serve as important signals of identity and affiliation. Hence, discursive deadlock might not result from a failure to access the reasons motivating opponents, so much as a disinclination to ally oneself with them (or even appear to be doing so). In that case, adopting alternative frames in one's speech may be altogether counterproductive, failing to address the real barrier to reasonable collective decision-making while incurring real social costs. We saw why this might be an especially pressing problem in social media, and how structural changes to platform design could potentially help foster more constructive debate. In this way, I believe a suitably enriched analysis of framing could open up new possibilities for improving political discourse in the digital age.

Acknowledgements

I am extremely grateful for helpful discussions about this material with participants at the UCL workshop on 'Social affinity and the uses of language,' the LOGOS-Cardiff University workshop on 'The pragmatics and epistemology of online communication,' and the UCL / KCL / NYU-Abu Dhabi workshop on 'Practical philosophy'. I am further indebted to two reviewers for this journal, whose insightful comments enabled me to make substantial improvements. The work was supported by UKRI (grant reference MR/V025600/1).

REFERENCES

- Aly, W. & Simpson, R. M. (2018) Political correctness gone viral. In C. Fox & J. Saunders (eds.) *Media ethics, free speech, and the requirements of democracy* (pp. 125-143). Routledge.
- Anderson L. and Lepore, E. (2013a). Slurring Words. *Noûs*, 47(1), 25-48.
- (2013b). What Did You Call Me? Slurs as Prohibited Words. *Analytic Philosophy*, 54(3), 350-363.
- Barnes, M. R. (2023). Who do you speak for? And how? Online Abuse as Collective Subordinating Speech Acts. *Journal of Ethics and Social Philosophy*, 25(2), 251-281.

- Beaver, D. and Stanley, J. (2023). *The politics of language*. Princeton University Press.
- Bermúdez, J. L. (2020). *Frame it again: New tools for rational decision-making*. Cambridge University Press.
- Bermúdez, J. L. (2022). Rational framing effects: A multidisciplinary case. *Behavioral and Brain Sciences*, 45, e220.
- Billingham, P. and Parr, T. (2019). Online public shaming: Virtues and vices. *Journal of Social Philosophy*, 51(3), 371-390.
- Bolinger, R. J. (2017). The Pragmatics of Slurs. *Noûs*, 51(3), 439-462.
- Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3), 297-326.
- Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 42, 419-450.
- Camp, E. (2019). Perspectives and frames in pursuit of ultimate understanding. In S. R. Grimm (ed.) *Varieties of understanding: New perspectives from philosophy, psychology, and theology* (pp. 17-46). Oxford University Press.
- Camp, E. and Flores, C. (2024). Playing with labels: Identity terms as tools for building agency. *Philosophical Quarterly*, 74(4), 1103-1136.
- Camp, E. and Nowak, E. (2025). Linguistic variation, agency, and style. In E. Lepore & U. Stojnic (eds.) *The Oxford Handbook of Contemporary Philosophy of Language* (pp. 676-712). Oxford University Press.
- Cepollaro, B. (2015). In defence of a presuppositional account of slurs. *Language Sciences*, 52, 36-45.
- Davies, A. (2025). Sharing Content Online: the Effects of Likes and Comments on Linguistic Interpretation. In P. Connolly, S. Goldberg, and J. Saul (eds.) *Conversations online: Explorations in philosophy of language* (pp. 64-82). Oxford University Press.
- Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87-100.
- Eckert, P. (2024). Social meaning. *Theoretical Linguistics*, 50(1-2), 63-70.
- Fisher, S. A. (2022a). Defining preferences over framed outcomes does not secure agents' rationality. *Behavioral and Brain Sciences* 45, E227.
- Fisher, S. A. (2022b). Frames, reasons, and rationality. *International Journal of Philosophical Studies* 30(2), 162-173.
- Frost-Arnold, K. (2021). The Epistemic Dangers of Context Collapse Online. In J. Lackey (ed.) *Applied Epistemology* (pp. 437-456). Oxford University Press.
- Goffman, E. (1969). *The presentation of self in everyday life*. Allen Lane.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.
- Goldberg, S. (2021). The promise and pitfalls of online 'conversations'. *Royal Institute of Philosophy Supplement*, 89: 177-193.
- Henderson, R. and McCready, E. (2024). *Signaling without saying: The semantics and pragmatics of dogwhistles*. Oxford University Press.
- Hom, C. (2008). The Semantics of Racial Epithets. *Journal of Philosophy*, 105, 416-440.
- Jeshion, R. (2013). Expressivism and the Offensiveness of Slurs. *Philosophical Perspectives*, 27, 231-259.
- Johnstone, B. (2017). Characterological figures and expressive style in the enregisterment of linguistic variety. In C. Montgomery & E. Moore (eds.) *Language and a sense of place* (pp. 280-300). Cambridge University Press.
- Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39(4), 341-350.

- Keiser, J. (2023). *Non-ideal foundations of language*. Routledge.
- Khoo, J. (2017). Code words in political discourse. *Philosophical Topics*, 45(2), 33-64.
- Khoo, J. (2021). Code words. In J. Khoo and R. K. Sterken (eds.) *The Routledge handbook of social and political philosophy of language* (pp. 147-160). Routledge.
- Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76(2), 149-188.
- Malinowski, B. (1923). The Problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards (eds.) *The Meaning of meaning: A study of the influence of language upon thought and of the science of symbolism* (pp. 296-336). Harcourt, Brace & World, Inc.
- Marwick, A. E. & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *new media and society*, 13(1), 114-133.
- Mason, L. (2018). *Uncivil agreement: How politics became our identity*. The University of Chicago Press.
- McDonald, L. (2021). Please like this paper. *Philosophy*, 96, 335-358.
- McDonald, L. (2025). Context collapse online. In P. Connolly, S. Goldberg, and J. Saul (eds.) *Conversations online: Explorations in philosophy of language* (pp. 83-105). Oxford University Press.
- Meyerowitz, J. (1986). *No sense of place: The impact of electronic media on social behavior*. Oxford University Press.
- Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2), 141-161.
- Nowak, E. (2022). Sociolinguistic Variation, Speech Acts, and Discursive Injustice. *Philosophical Quarterly*, 73, 1024-1045.
- Nowak, E. (2025). Sociolinguistic variation, slurs, and speech acts. *The Journal of Philosophy*, 122(7), 317-345.
- Nunberg, G. (2018). The social life of slurs. In D. W. Harris, D. Fogal, and M. Moss (eds.), *New Works on Speech Acts* pp. (237-295). Oxford University Press.
- Pinsof, D., Sears, D. O., and Haselton, M. G. (2023). Strange bedfellows: The alliance theory of political belief systems. *Psychological Inquiry*, 34(3), 139-160.
- Pullum, G. K. (2018). Slurs and obscenities: Lexicography, semantics, and philosophy. In D. Sosa (ed.) *Bad words: Philosophical perspectives on slurs* (pp. 168-192). Oxford University Press.
- Richard, M. (2008). *When Truth Gives Out*. Oxford University Press.
- Ronson, J. 2015. *So you've been publicly shamed*. Picador.
- Saul, J. (2018). Dogwhistles, Political Manipulation, and Philosophy of Language. In D. W. Harris, D. Fogal, and M. Moss (eds.), *New Works on Speech Acts* (pp. 360-383). Oxford University Press.
- Schlenker, P. (2007). Expressive presuppositions. *Theoretical Linguistics*, 33(2), 237-245.
- Stanley, J. (2015). *How propaganda works*. Princeton University Press.
- Tanesini, A. (2018). Intellectual servility and timidity. *Journal of Philosophical Research*, 43, 21-41.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, 59(4), S251-S278.

SARAH A. FISHER is a Lecturer in Philosophy at Cardiff University. She specialises in the philosophy of language, exploring the nature of meaning and communication. She is co-author with Emma Borg of *Meaning: A Very Short Introduction* published by Oxford University Press. She has published widely on meaning, framing, and digital speech. Her work is informed by empirical research in the cognitive sciences and addresses questions of social and political importance.

ADDRESS: Cardiff University College of Arts, Humanities and Social Sciences, John Percival Building, Colum Drive, Cardiff, UK, CF10 3EU. E-mail: fishers15@cardiff.ac.uk – ORCID: 0000-0003-1115-6134



HOW GENERAL ARE MARR'S LEVELS OF ANALYSIS? AN ASSESSMENT OF BERMÚDEZ'S VIEW

(¿Cómo de generales son los niveles de análisis de Marr?
Una evaluación de las propuestas de Bermúdez)

Marko Jurjako*

University of Rijeka

<https://orcid.org/0000-0002-7252-8627>

Keywords

Active inference
Free-energy principle
José Bermúdez
Interface problem
Personal/subpersonal explanations
Marr's levels of analysis

ABSTRACT: In his seminal book, *Philosophy of psychology: A contemporary introduction*, José L. Bermúdez argues that David Marr's levels of analysis cannot be used as a general framework for explaining phenomena in cognitive science. More specifically, he argues that Marr's levels of analysis paradigmatically apply to subpersonal modular processes, while the mind as a whole is also characterized by non-modular cognitive systems. In this paper, I evaluate Bermúdez's arguments for this conclusion. Part of the evaluation will be based on recent advancements in the active inference framework, suggesting that the mind as a whole can be analyzed within the Marrian paradigm. Moreover, I provide some reasons for thinking that Marr's levels of analysis could also be employed to illuminate the relationship between personal and subpersonal explanations.

Palabras clave

Inferencia activa
Principio de energía libre
José Bermúdez
Problema de la interfaz
Explicaciones personales/
subpersonales
Niveles de análisis de Marr

RESUMEN: En su obra seminal, *Philosophy of psychology: A contemporary introduction*, José L. Bermúdez argumenta que los niveles de análisis de David Marr no pueden ser usados como un marco general para explicar fenómenos en ciencias cognitivas. Más específicamente, argumenta que los niveles de análisis de Marr se aplican paradigmáticamente a procesos modulares subpersonales, mientras que la mente como un todo también está caracterizada por sistemas cognitivos no modulares. En este artículo, evalúo los argumentos de Bermúdez a favor de esta conclusión. Parte de esta evaluación se basará en avances recientes en el marco de inferencias activas, sugiriendo que la mente como un todo puede ser analizada dentro del paradigma de Marr. Además, proporcionaré razones para pensar que los niveles de análisis de Marr también pueden emplearse para iluminar la relación entre explicaciones personales y subpersonales.

Gako-hitzak

Inferentzia aktiboa
Energia askearen printzipioa
José Bermúdez
Interfazearen arazoa
Azalpen pertsonalak/azpipertsonalak
Marren analisi-mailak

LABURPEN: José L. Bermúdezek *Philosophy of Psychology: A Contemporary Introduction* lan aitzindarian dio David Marren analisi-mailak ezin direla marko orokor gisa erabili zientzia kognitiboko fenomenoak azaltzeko. Zehazkiago dio Marren analisi-mailak modu paradigmatikoa aplikatzen zaizkiela prozesu modular azpipertsonalei, baina oro har gogamenak sistema kognitibo ez-modularrak dituela. Lan honetan, ondorio horretara iristeko Bermúdezek emandako argudioak ebaluatuko ditut. Ebaluazio horren zati bat inferentzia aktiboaren esparruan duela gutxi izandako aurrerapenetan oinarrituko da, zeinen arabera, gogamena, oro har, Marren paradigmaren barruan azter baitaiteke. Gainera, zenbait arrazoi emango ditut frogatzeko Marren analisi-mailak azalpen pertsonalen eta azpipertsonalen arteko harremana argitzeko ere erabil litezkeela.

* **Correspondence to:** Marko Jurjako. University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy and Division of Cognitive Sciences, Sveučilišna avenija 4, 51000 Rijeka, Croatia. – mjurjako@ffri.uniri.hr – <https://orcid.org/0000-0002-7252-8627>

How to cite: Jurjako, Marko (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>).

Received: 11 January, 2024; Final version: 04 February, 2025.

ISSN 0495-4548 - eISSN 2171-679X / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

1. *Introduction*¹

Cognitive phenomena can be studied at different levels of analysis (see, e.g. Colombo & Knauff, 2020; Piccinini, 2020; Verdejo & Quesada, 2011). In this regard, David Marr (1982) provided an especially influential model for understanding cognitive phenomena (see, e.g. Peebles & Cooper, 2015; Poggio, 2012). He introduced the distinction between the computational, algorithmic, and implementational levels of analysis. Given its influence in cognitive science, many have critically discussed different aspects of Marr's levels of analysis and their relative importance (see, e.g., the papers in Peebles & Cooper, 2015; Zednik, 2018). Yet, few have explicitly discussed the capability of Marr's levels of analysis to provide a general framework for understanding psychological phenomena, including also personal and subpersonal perspectives on cognitive systems.

In a rare discussion of this issue, José Luiz Bermúdez (2005, p. 27) argues that Marr's distinction cannot be used for this purpose. He offers two related arguments to support this claim. First, he contends that Marr's distinction is best suited for explaining modular cognitive systems (see, also Cooper & Peebles, 2015). Marr's framework requires a clearly defined task and well-specified functions to generate hypotheses about a system's representations and algorithms. However, non-modular systems often lack such clear functional specifications, making them difficult to analyze using Marr's tripartite approach.

Second, Bermúdez argues that even if one successfully applies Marr's framework to a particular personal or non-modular cognitive capacity, it is very difficult—or even impossible—to extend this analysis to the mind as a whole. By this, he suggests that Marr's levels cannot provide a unified explanatory framework for understanding the general relationship between different levels of explanation in cognitive science, particularly between personal and subpersonal perspectives on the mind/brain.

The aim of this paper is to evaluate Bermúdez's claim that Marr's levels of analysis cannot provide a general framework for explaining cognitive processes. To address the first argument, I will show that cognitive (neuro)science offers computational models of higher-order, non-modular systems that can be usefully analyzed using Marr's tripartite distinction. To address the second argument, I will argue that Marr's framework can be applied to the mind as a whole and even shed light on the relationship between personal and subpersonal explanations, provided that a sufficiently rich theoretical framework is adopted. As a proof of concept, I will show that the active inference framework, grounded in the free-energy principle, provides a unified framework for understanding the overall function of the mind and its physical implementations. Given its computational structure, this framework lends itself well to analysis through Marr's levels.

The paper proceeds as follows. Section 2 summarizes Marr's levels of analysis. Section 3 contextualizes the debate on the generality of Marr's framework within Bermúdez's "interface problem". There I assess Bermúdez's first argument that Marr's levels of analysis are limited because they mostly apply to subpersonal and modular cognitive processes. Section 4, in a preliminary way, examines Bermúdez's second argument, which claims that Marr's levels of analysis, even if applicable to some higher-level non-modular processes, are inadequate for addressing the function of the mind as a whole. In Section 5, I defuse this argument by relying on the active inference framework, arguing that it provides sufficient conceptual resources for understanding the function of the mind as a whole. Section 6 revisits the interface problem, exploring how the application of Marrian analysis within the active inference framework can clarify the relationship between personal and subpersonal explanations.

2. *Marr's levels of analysis*

Marr (1982) famously argued that a complete understanding of some cognitive capacity involves an analysis at computational, algorithmic, and implementational levels. To illustrate how this framework is used, let us consider the human

¹ This paper builds upon and provides a more comprehensive elaboration of some of the ideas initially presented in Jurjako (2023).

ability to visually recognize objects. At the computational level, we are trying to determine what the system under consideration is doing and why, i.e., what is the function it is supposed to perform.² We can say that the function of visual object recognition is to recognize objects by identifying and classifying them based on their visual features, such as shape, size, color, and texture. In computational terms, this function of visual object recognition can be framed as a process in which the goal is to map the visual input (e.g., a 2D image on the retina) to a representation of a 3D object that is classified under some label (e.g., a cat or a dog).

At the algorithmic level, the aim is to determine how the system implements the functions and tasks identified at the computational level. This involves determining an algorithm that describes the steps and procedures that the system uses to solve the identified computational problem, and to identify the representations it uses to achieve this goal.³ In the case of visual object recognition, an algorithm for identifying and classifying 3D objects could use procedures for filtering visual inputs to detect edges of objects. As edges of objects are correlated with sudden changes in the intensity of the light that is reflected by an object, they can be represented by mathematical procedures that use zero-crossings. Zero-crossings are the points in an image where the intensity of the pixels changes from positive to negative values, or *vice versa*. Then, the algorithm may use procedures that extract features such as shape, size, and texture of an object from the zero-crossings. Finally, to classify the object, the algorithm may compare the extracted features with information about objects that is already stored in memory.

The implementational level aims to specify the neural structures and mechanisms that underpin algorithms involved in visual object recognition. For example, it has been proposed that the visual system uses a hierarchy of processing stages, with early stages dedicated to basic features such as edges, and later stages dedicated to more complex features such as shape, size, and texture. Some of the specific neural structures involved in this process are the retina, the primary visual cortex (V1), and the inferotemporal cortex (IT) (see, e.g. DiCarlo *et al.*, 2012).

Marr's levels of analysis were devised for analyzing cognition understood as an information processing system (Poggio, 2012). It is unclear whether this analysis can be used for understanding the general relation between different levels at which psychological phenomena can be explained, including the relation between the personal and the subpersonal. This question has been surprisingly underexplored in the literature. In a rare discussion of this issue, Bermúdez contends that Marr's levels of analysis are unsuitable for this task, as will be discussed in the next section.

3. *Can Marr's distinction be used as a general framework for analyzing phenomena in cognitive science?*

3.1. PERSONAL AND SUBPERSONAL EXPLANATIONS, AND THE INTERFACE PROBLEM

Bermúdez situates his arguments within the discussion of the relation between personal and subpersonal explanations (see Bermúdez, 2005, pp. 17-38). There is some uncertainty about how to clearly distinguish between the personal and the subpersonal (for recent discussions, see Dänzer, 2023; Drayson, 2012; Westfall, 2022). Nonetheless, for the purposes of our discussion, Bermúdez offers a reasonably clear approach to conceptualizing the distinction.

² For a seminal discussion of Marr's computational level, see Shagrir (2010); for a recent application, see Mallory (2024).

³ Marr is working in the classical paradigm of cognitive science that understands the mind based on the digital computer metaphor. Thus, it was natural for Marr to talk about representations and algorithms for transforming representations in a way that computing devices, such as cashier registers, manipulate representations. However, nothing in the idea of the algorithmic level as conceived by Marr commits one to adopt the idea of digital representations that are manipulated according to their structural features (see, e.g. Zednik, 2018). The minimal presupposition for the application of the Marrian paradigm is that cognitive processes might be interpreted as information processing devices, without specific commitments about how a cognitive system under examination represents information and what types of equations are used to describe their transformations.

In line with a venerable tradition in philosophical psychology, Bermúdez understands personal explanations as referring to traits that persons, or more broadly, rational agents, can properly possess (Dennett, 1969; Hornsby, 2000). Specifically, he claims that the personal is characterized by three features (see Bermúdez, 2000, p. 64). The first feature is its distinctive vocabulary, which refers to intentional states such as beliefs, desires, and intentions. The second feature is that personal explanations individuate a special class of regularities that might not be discernible from other perspectives. For instance, personal explanations often reference intentional actions, such as intending to make a money transfer, which can be realized in various physical ways—like making hand gestures, giving a verbal command, or other movements that share nothing in common except being instances of the same regularity recognized as “transferring the money” (see, also Fodor, 1974). The third feature of the personal level is that its explanations are guided by principles of rationality (Davidson, 2001; Dennett, 1981). For instance, in explaining why someone transferred money, we might assume that they believed the transfer would settle a debt and wanted to resolve it, thus choosing a method—like signing a check or making an electronic payment—that would achieve that goal (for influential discussions of rationality and its various applications across the social sciences and humanities, see Bermúdez, 2011, 2020).

Bermúdez also suggests that these features characterize the folk-psychological explanations we use in everyday life for social interaction.⁴ We typically understand people as possessing intentional states and assume that they act to satisfy their goals in light of their beliefs, in a minimally rational way. He further notes, however, that the personal level cannot be entirely equated with folk-psychological explanations, as the personal perspective is also employed in cognitive psychology. In this field, scientists measure reaction times, use verbal reports, and formulate hypotheses about conscious cognitive and perceptual abilities, thereby contributing to the development of scientific theories of personal-level states and processes (Bermúdez, 2000, p. 65). While it is reasonable not to equate personal explanations entirely with folk-psychological ones, we will set this distinction aside in the following discussion. This is because scientific psychology's personal-level constructs can be seen as more precise extensions of folk-psychological constructs and associated principles (see Bermúdez, 2005, p. 65). Moreover, insofar as the personal level is characterized by folk-psychological explanations and constructs, it becomes easier to frame the interface problem as Bermúdez does. However, a proper understanding of the interface problem requires a clear contrast between personal and subpersonal explanations.

The subpersonal is typically associated with states, activities, and processes that operate below the level of a person. Although subpersonal constructs are in some sense related to those we apply at the personal level, they should be clearly distinguished (see Bermúdez, 2005, ch. 2; Drayson, 2012). For example, Marr's theory of vision is commonly regarded as a subpersonal theory of visual processing. As explained in Section 2, according to Marr, the ultimate task of early vision is to produce a 3D representation of the external scene. However, the concept of representation in Marr's account should be clearly distinguished from the notion of representation of a visual scene we attribute to whole persons. The main difference is that unlike personal representations, which are typically consciously entertained, Marr's representational processes operate subconsciously.⁵ Moreover, Marr's theory of vision provides computational principles intended to explain how subconscious representations in early vision are processed to eventually yield a representation of the external environment that our cognitive system might further use for object recognition and classification. At the per-

⁴ Recent scholarship suggests that associating personal and folk-psychological explanations only with intentional states and rational principles is overly simplistic. Such explanations also include personality traits, habitual actions, social scripts, and references to skills that do not necessarily involve intentional states or rational constraints (Andrews *et al.*, 2021; Westfall, 2022; see, also Bermúdez, 2005, ch. 7). However, for the purposes of this paper and formulating the interface problem, Bermúdez's three-feature characterization of the personal and its relation to folk psychology will suffice.

⁵ It is worth noting that sometimes, the primary criterion for individuating the personal level involves references to conscious states and processes, while the subpersonal level is understood as pertaining to subconscious states and processes. However, although the personal perspective often refers to conscious states—and rightly so, since our conscious experiences permeate personal-level functioning (Chappell, 2023)—its explanations and associated states cannot be individuated solely by the criterion of consciousness. This is because there may be subconscious beliefs attributed to a person, and we may also explain people's behavior by referring, for instance, to character traits that do not necessarily operate at the conscious level (Westfall, 2022; see also the discussion in Bermúdez, 2005, pp. 29–31).

sonal level, however, we are concerned with reasoning processes that rely on conscious perceptions, which contribute to forming propositional attitudes and provide inputs to our decision-making processes characteristic of intentional agency.

Subpersonal explanations are also linked to important distinctions in how different disciplines approach the study of the mind (Jurjako, 2024). The personal perspective is used when we consider mental states, processes, and abilities attributed to people understood as agents capable of rational thought, while the subpersonal perspective is typically applied in cognitive science and neuroscience, where people's perceptual and cognitive abilities are understood as products of mechanisms, where they can be explained at different levels of analysis, involving their functions, causal roles, and structural composition (Shea, 2018). In this respect, the subpersonal level can be characterized by various levels of analysis typically studied in cognitive (neuro)science. These include cognitive-computational processes as explored in cognitive psychology, analyses of how populations of neurons enable cognitive and motor capacities, detailed explanations of individual neuronal responses to specific stimuli, and even more fine-grained aspects studied in molecular neuroscience. Each of these levels contributes to our understanding of how subpersonal processes underlie complex cognitive and behavioral phenomena (see Bermúdez, 2022).

According to Bermúdez, considering the personal and various levels of subpersonal explanation leads to the question of how to understand their relationship. This is what he calls the interface problem:

This is the problem of explaining the relation between the commonsense, everyday type of psychological explanation that we all engage in every day [...] and the levels of explanation lower down in the hierarchy. (Bermúdez, 2005, p. 35)

To further illustrate the interface problem, he articulates it through a series of questions:

How do explanations of the behavior of people given in terms of their beliefs, desires and other psychological states mesh, for example, with explanations in terms of patterns of activity across populations of neurons? How does the biochemistry of what goes on inside a neuron relate to the dynamics of how a person interacts with the environment? What is the relation between understanding a person as a conscious, reasoning agent, on the one hand, and understanding that person's brain as a complicated type of computational mechanism? (Bermúdez, 2005, p. 35)

As we can see here, Bermúdez formulates the interface problem as one of understanding a hierarchy of explanations, where personal explanations occupy the upper levels of the hierarchy and subpersonal explanations capture the lower levels. According to his view, one of the main tasks of the philosophy of psychology is to provide a coherent model for understanding the relationship between these levels (Bermúdez, 2005, p. 35). Given that Marr's levels of analysis offer a clear approach to understanding and explaining cognitive phenomena, it is worth considering whether this framework can sufficiently illuminate the relationship between these levels of explanation.

At first glance, it seems that the interface problem could be straightforwardly addressed from a Marrian perspective.⁶ Marr's levels of analysis —delineated by the questions “why”, “how”, and “where”— can be seen as aligning with the levels of explanation Bermúdez discusses. In this view, the personal level, which largely involves folk-psychological explanations, could be understood as providing explananda with functions specified at the computational level. The algorithmic level would then seek to identify the algorithms and representations that give rise to these psychological functions. Finally, the implementational level would investigate their biological underpinnings.

⁶ It is worth noting that an often-overlooked question is whether the interface problem should be addressed uniformly across cognitive science or if different fields require distinct solutions. I touch on this issue in a recent philosophy of psychiatry paper, where I discuss the notion of mental disorder in relation to the interface problem (see Jurjako, 2024).

However, Bermúdez offers reasons for doubting that this straightforward solution can work. In the following sections, we will examine the cogency of his skepticism regarding the use of Marr's levels of analysis for addressing the interface problem.⁷

3.2. MODULAR AND NON-MODULAR SYSTEMS, AND MARR'S LEVELS OF ANALYSIS

One reason for thinking that Marr's levels of analysis are particularly limited in explaining cognitive phenomena is that they only pertain to modular systems. In this regard, Bermúdez writes that "it should be clear that nothing like Marr's account could be *straightforwardly* applied to what we might think of as higher (i.e. non-modular) cognitive processes" (Bermúdez, 2005, p. 27). The reasoning seems to be as follows. Marr's levels of analysis are well suited for explaining modular systems. Here Bermúdez (2005, p. 25) employs Jerry Fodor's classical (1983) notion of modularity.⁸ According to this view, modular systems are, among other things, domain specific (they are specialized for well-specified tasks) and informationally encapsulated (they only respond to a finite set of inputs and cannot be penetrated with other types of information). This implies that modular systems possess clear functional specialization with evident implementability through well-defined algorithms. In contrast, personal states are typically associated with non-modular processes. This apparently means that they will lack clearly specified functions that can be algorithmically implemented. Thus, typically it would not be feasible to apply Marr's analysis to personal non-modular processes.

Even though Marr's levels of analysis well capture modular processes, it does not follow that they are not useful for explaining non-modular processes. Consider, for instance, our ability for ordinary reasoning. This ability is informationally "promiscuous", in the sense that it can take in information from different cognitive systems, integrate it, and make further inferences based on it. Moreover, such an ability is not domain-specific, at least not in the sense that it is restricted to responding to a narrow set of inputs. For example, a person can by seeing the weather outside form the belief that it is sunny outside and by hearing the rain form the belief that it is raining outside. By integrating these beliefs, the person can conclude that it is sunny and raining at the same time.

More importantly, such a non-modular ability can be usefully analyzed at the computational, algorithmic, and implementational levels. For example, at the computational level, this ability can be understood as the ability for reasoning that takes symbolic representations as input and outputs other symbolic representations. At the algorithmic level, we can think about different representational systems and associated rules for their processing. For instance, we can formulate algorithmic procedures based on Bayesian statistical inferential procedures or inferential procedures based on some form of defeasible logic. Finally, we can consider where and how these logical procedures are implemented in the brain (see, e.g. Baggio *et al.*, 2015). Thus, apparently non-modular systems can have well-specified functions that are amenable to Marrian analysis.

Against such a piecemeal response, Bermúdez offers what I take to be his second argument for the limitations of Marrian analysis. This argument takes the form of a more general objection.⁹ He asserts that

⁷ In what follows, I will refer to Bermúdez's argument as presented in his (2005) book. The same arguments are also discussed in his cognitive science textbook (see 2014, pp. 126-129). Interestingly, in the last (fourth) edition of the textbook, the discussion of these arguments has been largely omitted.

⁸ For the sake of discussion, I follow Bermúdez in assuming the Fodorian view of modularity. There are other non-Fodorian perspectives on how modular processes could be understood. In the next section, I will introduce the active inference framework, which is often understood as challenging the sharp distinction between modular and non-modular systems (for discussion, see Drayson, 2017). Given that this issue is not directly pertinent for the present discussion, it will not be further examined.

⁹ In addition to the following objection, Bermúdez (2005, pp. 26-27) also suggests that Marr's levels of analysis might be limited to modular processes because algorithmically analyzing non-modular processes might face the frame problem—a challenge in cognitive science and AI where a system struggles to determine which aspects of an ever-changing environment are relevant to its current goal, making it difficult to algorithmically handle complex, context-sensitive tasks. However, this objection does not seem too serious. While Bermúdez is right that the frame problem poses a challenge for creating algorithms for non-modular processes, this is not a principled problem for applying the Marrian analysis because its usefulness does not depend on whether we can at this moment invent algorithmic procedures that would be

[E]ven if [Marr's levels of analysis] could be extended to non-modular processes, [...] it will certainly be *impossible* to do so for the mind as a whole—and it is, of course, an understanding of the mind as a whole that we are ultimately aiming for. Marr's analysis of the early visual system provides a clear illustration of the general idea of a hierarchy of different levels of explanation. But it is not itself pitched at the right sort of level to provide a model of how we might understand the general idea of a hierarchy of explanation applied to the mind as a whole. (Bermúdez, 2005, p. 27, emphasis added)

There appear to be two distinct issues in Bermúdez's quote. The first relates to the claim that Marr's levels of analysis cannot be applied to the mind as a whole. We will explore Bermúdez's reasoning for this in the next section. The second issue pertains to Bermúdez's assertion that Marr's levels are unsuitable as a model for understanding "the general idea of a hierarchy of explanation applied to the mind as a whole". These two claims seem connected as a premise (first claim) leading to a conclusion (second claim). Specifically, if we ask why Bermúdez claims that Marr's levels of analysis do not provide a suitable framework for explaining the mind as a whole, the answer seems to be that this is because the mind, as an integrated system, cannot be fully analyzed through Marr's levels.

The more interesting claim seems to be the first one, and later on I will focus on it. This is because even if Bermúdez is correct that Marr's levels of analysis do not offer a clear model of the hierarchy of explanations in cognitive science, this would not be particularly surprising or controversial. Indeed, there is general agreement among philosophers of cognitive science that Marr primarily proposed levels of *analysis* of cognitive phenomena, without specifying how these levels should align with different types of explanations applied to cognitive phenomena. To address this gap, several philosophers have suggested connecting Marr's levels of analysis with mechanistic explanations (see, e.g. Bechtel & Shagrir, 2015; Zednik, 2018). According to such approaches, Marr's levels of analysis can naturally be understood as applying to mechanisms underlying cognitive phenomena of interest by guiding us to ask questions about a mechanism's function, how it performs that function, and where it is physically implemented. Given that mechanisms are composed of parts and processes that can be structured at different levels, interpreting Marr's proposal within a mechanistic philosophy of science offers a clear view of how to formulate a Marrian-inspired model of explanatory hierarchy within cognitive science (for further discussion, see Chirimuuta, 2024; Zednik, 2018).

However, if Bermúdez is correct that Marr's levels of analysis cannot be applied to the mind as a whole, then it would follow that no matter how detailed our Marrian mechanistic explanations are, they would be lacking in certain respect, and would not enable us to provide a coherent framework for thinking about cognitive phenomena spanning different levels of description. For this reason, in what follows, I will focus on evaluating Bermúdez's claim that Marr's levels of analysis cannot be extended to characterizations of the mind as a whole.¹⁰

4. *Why Marr's levels of analysis cannot be applied to the mind as a whole?*

It is not immediately clear why Bermúdez thinks that a Marrian analysis cannot be applied to the mind as a whole. One reason that emerges in Bermúdez's discussion relates to the claim that Marr's analysis cannot be applied to people as they are understood at the personal level of functioning. This line of reasoning begins with a reasonable claim made by Bermúdez that

able to model every aspect of the mind. Theorizing and practical work in cognitive science starts with the presupposition that the human mind has somehow solved the frame problem. And by studying the mind from the computational, algorithmic, and implementation levels, we are trying to discover how our minds are capable of non-modular higher-level thinking and where these processes are implemented in physical systems such as the brain.

¹⁰ Thanks to an anonymous reviewer for pressing me to more adequately delineate the scope of this discussion.

Cognition is not an isolated activity and if we are interested in studying the mind as a whole we must start from the twin facts, first, that it is organisms that have minds and, second, that possessing a mind allows those organisms to behave in the ways characteristic of intelligent agents. (Bermúdez, 2005, p. 28)

The idea seems to be that to explain and understand the role of the mind as a whole, we must pay heed to its role in regulating the behavior of cognitive agents. People are paradigmatic cognitive agents. Thus, the explanation of the mind must begin with an explanation of the behavior of the person. However, according to Bermúdez:

Theories such as Marr's operate at a lower level than the level of cognitive agents. They deal with parts or modules of the cognitive agent, rather than with the agent itself as a thinking and acting organism. They are theories at the subpersonal level (below the level of the person). (Bermúdez, 2005, p. 28)

While it is true that Marr's specific theory of early visual processing is best understood as a subpersonal theory, this does not imply that his framework for levels of analysis cannot be extended to other, non-subpersonal cognitive phenomena. Indeed, as argued earlier, it seems we can apply Marr's levels of analysis to a paradigmatically non-modular personal-level ability, such as the ability for making ordinary/logical inferences. Why, then, should we assume that Marr's levels of analysis are limited in this regard?¹¹

There seems to be something about the concept of the mind as a whole that Bermúdez believes resists analysis in Marrian terms. The claim seems to be that if the mind as a whole cannot be provided a functional specification, then attempting a computational analysis would be futile and we would not be able to formulate explanations of how the mind actually performs these tasks or where these processes are implemented. Indeed, in this regard Bermúdez writes:

[W]hen we are thinking about the mind as a whole, there are difficulties applying the type of functional analysis that Marr applied to the early visual system. When we are thinking about the mind as a whole it is very difficult, and perhaps even impossible, to identify tasks that can be understood in a determinate enough way to yield algorithms. (Bermúdez, 2005, p. 28)

Essentially, Bermúdez seems to argue that using Marr's levels of analysis to provide a general framework for explaining cognitive phenomena spanning across personal and subpersonal levels is virtually impossible because the mind, as an integrated system, lacks a clear functional specification.

It is not entirely clear why Bermúdez believed that the mind as a whole lacks functional specification. Perhaps, at the time of his writing, philosophers of cognitive science struggled to conceptualize how non-modular cognitive systems, with their domain generality and cognitive penetrability, could be computationally analyzed.¹² Be that as it may, I will not focus on exploring additional reasons for this view. Instead, I will shift my approach and offer a proof of concept. Specifically, I will argue that there are no fundamental obstacles to using Marr's levels of analysis to understand the mind as a set of abilities characterizing the behavior of whole organisms—at least, not if one adopts a sufficiently ambitious framework for conceptualizing the mind. Additionally, I will argue that working within a sufficiently rich cognitive paradigm can shed light on the relationship between personal and subpersonal levels of functioning. To support this, I will introduce the active inference framework in the next section and illustrate how Marrian analysis can be applied within this paradigm to enhance our understanding of the mind as a whole.

¹¹ Richard Cooper and David Peebles (2015, p. 253) also notice that Marr was primarily applying levels of analysis to components and sub-components of cognitive abilities, and not to whole agents. However, despite this observation, they do not claim that Marr's levels cannot in principle be applied to whole agents.

¹² In this regard, Bermúdez might have been influenced by Fodor's (1983) argument that there could never be a cognitive science of non-modular systems. Such systems are characterized by domain generality and holism, where any piece of information can influence any other cognitive process, leading to intractability (for discussion, see Murphy, 2019, see also footnote 8 above).

5. Active inference and the mind as a whole

5.1. THE ACTIVE INFERENCE FRAMEWORK¹³

The active inference framework is based on the Free Energy Principle (FEP). According to the FEP, the general function of the mind can be understood as minimizing free energy (for an overview, see, e.g. Mann *et al.*, 2022).¹⁴ More generally, the FEP prescribes the general conditions for self-organizing systems to remain in existence. The survival of all self-organizing systems, such as living organisms, depends on their ability to resist the tendency to dissipate under the environmental pressures. The FEP states that self-organizing systems maintain their integrity and resist dissipation by minimizing atypical or surprising events in their environment. Granting the insight that organisms possess minds that enable intelligent behaviors, it follows in the present context that we can construe the general function of the mind as the ongoing process of minimizing uncertainty, which will later be expounded in terms of free energy minimization.¹⁵

To better understand the basic notions underpinning the FEP, we need to elaborate on what is meant by surprise, self-organizing systems, and the role of free-energy in minimizing surprisal. We will start with the formal notion of surprisal.

In information theory, surprise is a measure of the amount of unexpectedness associated with a particular event. To distinguish it from the commonsensical notion of surprise, researchers tend to call this quantity surprisal. It is defined as the negative logarithm of the probability of an event occurring:

$$\text{Surprisal: } -\log P(s) \quad (1)$$

Here, $P(s)$ denotes probability of s , $-\log$ is a logarithmic function of the probability function. Surprisal is inversely proportional to the probability of that event occurring. In other words, the higher the probability of an event, the lower the surprisal, and *vice versa*.

The general idea is that adaptive action involves remaining in familiar states by minimizing surprisal or uncertainty (see, e.g. Buckley *et al.*, 2017). For instance, for a fish, familiar states with low surprisal include being in water, while being outside water involves being in states with high surprisal. In the basic case, the surprisal of a state is determined by how expected it is for an organism to be in that state from the perspective of its phenotype and the ecological niche it inhabits. Thus, in this context, surprisal is determined by the conditional probability that an organism will experience an event or be in some state given its phenotype. From a formal perspective, we can think about an organism's phenotype

¹³ In what follows, I provide a basic introduction to the formal machinery of active inference and the Free Energy Principle. An informed reader familiar with this literature may skip ahead to Subsection 5.2.

¹⁴ Some authors differentiate between high and low road approaches when introducing the active inference framework (see Parr *et al.*, 2022 chs. 2-3). The low road approach involves tracing the development of active inference from Helmholtz's idea of perception as inference all the way to contemporary predictive processing accounts, according to which the brain promotes adaptive behavior by minimization of prediction errors (for a notable book-length discussion, see Hohwy, 2013). In contrast, the high road approach involves the normative perspective, starting with first principles concerning the necessary conditions for organisms to maintain their existence, which in this case involves minimization of free energy. For the present purposes, introducing the framework from the high-road perspective seems preferable as it will enable us to more perspicuously think about the function of the mind as a whole.

¹⁵ An anonymous reviewer for this journal has rightly suggested that the claim in this section —that the framework based on the FEP provides a coherent way to conceptualize the mind as a whole— can be connected to the work of Julian Kiverstein and Matt Sims (2021). They argue that FEP can offer a “mark of the cognitive”, providing criteria to distinguish organisms whose behavior can be explained in cognitive, rather than non-cognitive, terms. While this connection is interesting and warrants further exploration, it lies beyond the scope of this paper. Our focus here is on whether Marrian analysis can be applied to the mind as a whole, rather than on distinguishing organisms by attributing cognitive processes.

as embodying a model of its relations with the environment it is embedded in. This leads us to the idea of a Markov blanket which determines the boundaries between self-organizing systems and their surrounding environment.

According to the FEP, Markov blankets determine the identity of a self-organizing system. A Markov blanket is a statistical construct originally introduced to capture efficient and qualitative forms of probabilistic inference (Pearl, 1988). The goal was to capture relevance and dependency relations among variables via probabilistic graphical models. In general, this is accomplished by identifying the set of variables that constitute the Markov blanket of the target variable. The Markov blanket includes the parents (direct causes) of the target variable, the children (direct effects) of the target variable, and any other variables that are probabilistically connected to the target variable, while excluding the conditionally independent variables (i.e. those that do not enable inferring the target variable), and thus ensuring that all relevant information and dependencies are captured.

In the context of FEP, Markov blankets are used to formalize the partitioning of a self-organizing system into its internal, external (environmental), and the boundary or blanket states that separate them (see, e.g. Hipólito *et al.*, 2021). The idea is that the internal states are separated, i.e. conditionally independent from the external states given the boundary states. The boundary states are further divided into sensory and active states. The relation of dependence is such that external states can influence sensory states that further cause changes in internal states. In contrast, the internal states can cause changes in active states that in turn cause changes in the external states.¹⁶

These relations form the core of the *active inference* framework. Sensory states represent perceptual processes that affect an organism's internal states, while active states play the role of actions that unidirectionally stem from internal states and cause changes in the external state. Together, the functioning of sensory and active states form causal perception-action loops that enable the organism and its internal states to efficiently adapt to the continuously changing environment and in this way to minimize surprisal. To adapt to the external environment an organism needs to be able to detect the external states and accordingly adjust its internal states or adjust the external states to make them conducive to its well-being. Adjusting internal states is called perceptual inference, while adjusting external states via action is called active inference.

Given the existence of a Markov blanket, surprisal can be minimized only indirectly, via the influence of sensory and active states because only through them can the internal and external states exchange energy, matter, and information (Hipólito *et al.*, 2021). In probabilistic terms, this means that the organism faces the problem of determining the posterior probability of external states given its blanket states. Formally, problems of this type are solved by employing Bayesian inference. To implement computational procedures that produce results approximating Bayesian inference, it has to be supposed that the brain (implicitly) embodies a generative model of how the external states causally affect its sensory states.¹⁷

The role of generative models is to represent the generative processes, that is, the causal structure of the external states and how they affect sensory states. In formal terms, a generative model is a joint probability distribution of two variables:

¹⁶ It should be noted that biological systems at different levels of spatiotemporal organization can be construed as having a Markov blanket (Hipólito *et al.*, 2021). For instance, the membrane of a cell can be conceived as a boundary of the cell's Markov blanket, without which the cell would dissipate and fall out of existence. Similarly, in the brain neurons can be understood as entities with Markov blankets, but also populations of neurons, whole brains, and entire organisms. Given this flexibility about which entities can be construed as having Markov blankets, and therefore fall under the purview of active inference, some have criticized the whole framework as not being sufficiently restrictive about delineating cognitive systems (for discussion, see Bruineberg *et al.*, 2022). I will not delve into the ontological disputes of the active inference framework. For my purposes, it is sufficient to think of Markov blankets as characterizing entire organisms and the neurobiological processes that underpin the relevant mental capacities.

¹⁷ It should be added that the debate is still ongoing whether the discourse about generative models and probabilistic Bayesian beliefs should be understood in realist or instrumentalist terms (for discussion, see Kirchhoff *et al.*, 2022; Kiverstein & Sims, 2021; Ramstead *et al.*, 2020). As this does not directly affect the discussion in the paper, I will remain neutral about this issue.

$$P(h, s) = P(s | h) * P(h) \quad (2)$$

Here, h is a variable of the internal model that represents a hypothesis about how sensory data s are generated by external processes. The generative model can be decomposed into the prior probability of h and the likelihood that s would be observed under the hypothesis h . Prior probabilities and likelihoods are often called beliefs that the brain has about the external generative processes.¹⁸ Given this generative model, the activities of the brain can be cast as implicitly engaging in inferential processes that compute the posterior probability of the hypothesis via Bayes theorem:

$$P(h | s) = (P(s | h) * P(h)) / P(s) \quad (3)$$

Importantly for our context, it should be noted that performing Bayesian inference involves maximizing model fit with the available evidence or data. Moreover, maximizing model fit (i.e. $P(s)$) is equivalent to minimizing surprisal of the model (i.e. $-\log P(s)$). From this follows that Bayesian inference is equivalent to information processing that minimizes surprisal of s (see Parr *et al.*, 2022, pp. 18-19).

The surprisal of the brain's sensory states can be minimized by Bayesian inference (or alternatively, its model evidence can be maximized) in two ways. The first is by performing perceptual inference where surprisal is minimized by sensory states. For instance, a person's generative model may ground an expectation (i.e. a prior belief) that a glass of water is on the table, but the incoming sensory data do not confirm this expectation. To reduce this surprisal, the person can update their generative model that the glass is not on the table. Alternatively, active states can reduce surprisal. For instance, the person could reduce surprisal by putting the glass of water on the table and thus changing the world instead of their sensory states.

However, estimating surprisal and minimizing it by Bayesian inference is often computationally intractable (see Parr *et al.*, 2022, p. 27). Surprisal depends on marginal probabilities of evidence and the more complex generative models get, the more variables need to be marginalized out to calculate the probability of evidence, which makes Bayesian inference intractable. This problem can be solved by using variational procedures that approximate optimal Bayesian inference. This is where free energy minimization becomes relevant.

According to FEP, the quantity that organisms actually minimize is variational free energy. In information theory, free energy provides an upper bound to an event's surprisal. Given that free energy is more feasible to estimate than surprisal, the basic claim of the active inference framework is that organisms minimize state uncertainty (i.e. their surprisal) by minimizing variational free energy. Formally, this means that posterior beliefs are replaced by a distribution that approximates them, called approximation posterior Q , while log probability of evidence is replaced by variational free energy F . Thus, optimal Bayesian inference is formulated as minimization of variational free energy F . Free energy is a functional $F[Q, s]$ that takes as arguments approximate posterior Q and sensory signals s (that represent available evidence for the generative model). There are several equivalent mathematical formulations of the free energy minimization principle (see Parr *et al.*, 2022, pp. 28-30). We will focus on the one that makes clear how minimizing free energy approximates minimization of surprisal:

$$F[Q(h), s] = DKL[Q(h) || P(h|s)] - \log P(s) \quad (4)$$

The first term is the Kullback-Leibler (KL) divergence between the approximate posterior distribution Q of a hypothesis h and the true posterior probability of the hypothesis given the sensory input (i.e. $P(h|s)$). The KL divergence meas-

¹⁸ This talk about "beliefs" should not be confused with beliefs as personal level states (see Dewhurst, 2017). Here, "beliefs" denote probability distributions that compose the generative models, and as such are typically taken to characterize the brain's subpersonal states and processes.

ures how different these two distributions are from each other. Minimizing this term means making $Q(b)$ as similar as possible to the true posterior $P(b|s)$. Minimization of KL is usually construed as involving perceptual inference. If perceptual inference reduces the KL divergence to zero, free energy would be equal to surprisal (i.e. $-\log P(s)$). The second term in the equation is the already familiar negative log probability of sensory evidence s that measures the surprisal of model evidence, i.e. how well the model's predictions match the observed data. Minimizing this term means maximizing the likelihood of the observed data given the model's predictions. This can be directly achieved only via active inference, i.e. by performing actions that affect which sensory evidence is available to the agent. In a nutshell, given that free energy F is an upper bound to surprisal, equation (4) shows that by minimizing the divergence between the approximate posterior $Q(b)$ and the true posterior $P(b|s)$ (via perceptual or active inference), the free energy of a generative model gets constantly closer to its surprisal (i.e. $-\log P(s)$).

In summary, according to the active inference framework, action and perception present two sides of the same coin, whose ultimate function is to minimize free energy of a generative model. In more familiar philosophical terminology, perceptual inference exemplifies a mind-to-world direction of fit (i.e., the brain adjusts its internal models to match the external reality it perceives), while active inference can be associated with a world-to-mind direction of fit (i.e., by initiating action that brings the environment in line with brain's internal expectations). With these considerations in view, in the next subsection, we turn to the task of explaining how within this framework Marr's levels of analysis can be fruitfully used to analyze the mind as a whole.

5.2. MARR'S LEVELS OF ANALYSIS AND THE ACTIVE INFERENCE FRAMEWORK

5.2.1. The computational level

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, emphasis in the original text). For organisms to survive and remain in homeostatic states they need to minimize surprisal, which is accomplished by minimizing free energy. The crucial component for achieving this is the supposition that the organisms' brains possess a "*generative model* that describes the problem the brain is trying to solve" (Parr *et al.*, 2022, p. 105, emphasis in the original text). Once this model is identified, by performing active inference the researcher can derive predictions about the behaviors, inferences, and associated neural dynamics, and compare the model with the attained data.

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.*, 2017). 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, Ramstead, *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.¹⁹

¹⁹ In the next section, I provide further considerations on how active inference can be employed to elucidate the relationship between the personal and subpersonal levels.

5.2.2. The algorithmic level

At the algorithmic level of analysis, this framework provides insight by offering process level theories of cognitive phenomena (Friston *et al.*, 2017). 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). The development of process-level theories involves elaborating on the specifics of a generative model that is designed to capture the cognitive or behavioral ability under investigation. This is achieved by considering several factors (see Parr *et al.*, 2022, pp. 106-113):

1. Markov blanket: firstly, we need to decide which system we are modelling. This is done by determining the Markov blanket of the system under investigation, including the determination of the internal, external, and the interfacing blanket states. For instance, the system might be the organism as a whole, but also its parts, such as the brain, brain parts, or other bodily or extra body regulatory systems that might be construed as performing inferences by minimizing free energy (see, e.g. Hipólito *et al.*, 2021; Parr *et al.*, 2022, p. 109).
2. Representational primitives: secondly, we need to choose the form of the generative model. This includes deciding on the appropriate type of variables, parameters, and the depth of the spatio-temporal hierarchy over which inferences and learning are defined. For instance, modelling perceptual processes that involve luminance contrasts would involve continuous variables, while modelling processes underlying object recognition would involve categorical variables.
3. Components of the generative model. Thirdly, we need to decide which parts of the model are fixed and which need to be learned. This includes determining the changeable parts of the model that are determined by the data, and priors that may be supplied by the researcher. Moreover, depending on the type of learning we are modelling, there will be a distinction between variables whose values are updated on faster timescales (e.g. perception of basic visual features of an object), and parameters whose values are updated on slower timescales (e.g. perception and recognition of whole objects).
4. Components of the generative process: finally, we need to determine the components of the generative *process* that the brain/mind is trying to infer based on its generative *model*, and how the generative process relates or is different from the elements composing the generative model.

By considering these factors we can formulate algorithmic theories of different cognitive capacities.

For instance, let us consider human decision-making abilities that enable choosing among the available options. Our Markov blanket of a person whose capacities we are modelling will determine the internal and external states, and their boundaries. Decision-making capacities and behavioral tasks that are used to elicit them are often modeled by partially observable Markov decision processes (POMDP) (see Smith, Friston, *et al.*, 2022). The structure of POMDP presupposes discrete time steps and state variables. Consequently, such models entail that beliefs composing generative models will be represented with categorical variables and parameters. Given such POMDP, they will involve beliefs about observational outcomes, states, actions, and policies. Moreover, depending on the behavioral task being modeled, a generative model can involve different hierarchies of beliefs whose function is to successfully predict sensory observations and enjoy actions that produce preferred outcomes.²⁰

²⁰ It is worth noting that process theories of perceptual and active inference are often based on the predictive coding schemes that involve continuous variables and are thought to provide a biologically plausible account of how the brain implements the FEP (see, e.g. Buckley *et al.*, 2017, p. 58). According to predictive coding, perception, learning, and action involve models that minimize prediction errors. Moreover, predictive coding serves as the foundation for the development of more comprehensive predictive processing theories of the mind/brain (Clark, 2016; Hohwy, 2013). In this regard, some consider predictive coding as more than a mere process theory that also offers a computational-level explanation of the brain (Sprevak & Smith, 2023). For more on the relation between predictive processing accounts and the active inference framework, see Hohwy (2013, chs. 2 and 4) and Sprevak and Smith (2023).

5.2.3. The level of implementation

At the implementational level, active inference allows for studying the brain at different levels of granularity. At the level of brain mechanisms, the active inference framework accounts for many predictive loops that characterize different brain systems and its general capacity for regulating internal processes and external environments (see Pezzulo *et al.*, 2022). From a finer-grained perspective, specific components of generative models underlying perceptual and active inferences are related with specific types of neuron populations and different aspects of neural functioning (for a recent experimental study, see Isomura *et al.*, 2023). For instance, studies of the implementations of priors and likelihoods of hierarchical perceptual models typically associate them with the functioning of deep and superficial pyramidal neurons, while learning and active inference are typically associated with the activity of dopaminergic neurons (for a short overview and references, see Smith, Friston, *et al.*, 2022, section 5).

In summary, according to the active inference framework, perception, learning, and action are conceptualized as distinct abilities that optimize the same function: the minimization of surprisal by reducing free energy. The generality of the active inference framework allows us to speculate about the holistic function of the mind and its implementation within various neural and bodily processes. Crucially, the generality of the framework also allows us to appreciate the value and feasibility of employing Marr's levels of analysis to comprehensively study the mind.

6. *Limits of active inference and the Marrian paradigm*

The active inference framework is very ambitious, with its proponents holding very optimistic views regarding its explanatory capabilities. To appropriate a quote from Jakob Hohwy, we might say that this framework purports to

[G]iv[e] the organizing principle for brain function as such. It should then encompass and illuminate all aspects of perception and action, including aspects that cognitive science and philosophy of mind view as problematic or poorly understood. (Hohwy, 2013, p. 101)

Naturally, due to its broad ambitions, various facets of the framework have been criticized (see, e.g. Bruineberg *et al.*, 2022; Colombo & Wright, 2017, 2018). In this regard, those who are skeptical of the scope or general plausibility of the framework might question its implications for our discussion of Marr's levels of analysis. In what follows, I explain why I think this skepticism does not significantly impact the direction of the present discussion.²¹

The criticisms span a spectrum, with two extremes concerning the generality and empirical adequacy of the framework. On one end, the FEP is often presented as a mathematical principle intended to provide an *a priori* account of formal conditions for modelling self-organizing systems (see Andrews, 2021). When FEP is understood in this way, it is not empirically falsifiable. Moreover, under this interpretation, applying Marr's levels of analysis would need to be reconsidered, because if the FEP lacked empirical constraints for examining mental processes, it would fail to meet the conditions required for analysis at Marr's algorithmic and implementational levels.

However, as discussed in the previous section, even though FEP can be understood as providing a high-level, computational task analysis applicable to all self-organizing agents, its empirical value lies in guiding the development of process theories that aim to model mechanisms, or at least provide sketches of mechanisms underlying various cognitive abilities. This is achieved by constructing generative models with different types of variables and parameters that can be

²¹ Thanks to an anonymous reviewer for prompting me to address more explicitly the role of active inference in examining the scope of Marr's levels of analysis.

used to test and account for the available cognitive and behavioral data. However, understanding FEP and active inference in more substantial terms leads to another set of concerns.

When active inference and FEP are understood as offering an empirical theory of biological self-organization or mind/brain function, concerns can be raised about its empirical adequacy. For instance, Matteo Colombo and Cory Wright (2017) argue that mental processes involving reward and motivation, which are underpinned by dopaminergic neurotransmitters, are better explained through a pluralistic approach that incorporates diverse theories rather than relying on a single unifying framework. In this view, FEP-based theories that interpret cognitive mechanisms as involving the minimization of free energy risk oversimplification, by blurring important distinctions across different biological systems, and making it difficult to validate specific mechanisms underlying adaptive behavior.

Such criticisms of the active inference framework, while important, are less relevant to the present discussion. Our focus is on whether Marr's levels of analysis can serve as a framework for understanding the mind as a whole. The purpose of using the FEP and active inference as a theory of the mind/brain is to illustrate (1) what a scientifically influential account of the mind as a whole might look like, and (2) how such an account can be meaningfully analyzed from a Marrian perspective. Admittedly, active inference may not ultimately provide a sufficiently adequate account to explain all cognitive phenomena and every aspect of brain function. However, considering such a general account of cognitive/brain function allows us to explore in a straightforward way what would be required to view the mind as a regulator of cognitive agents and to apply Marr's levels of analysis to the mind understood in this way. Given these considerations, I maintain that Bermúdez has not provided sufficient grounds for thinking that the Marrian framework cannot be applied to the mind as a whole.

6.1. REVISITING THE INTERFACE PROBLEM WITHIN ACTIVE INFERENCE

Nonetheless, it might still appear that the initial problem Bermúdez pointed out has not been resolved. Recall that Bermúdez (2005, ch. 2) discusses the limitations of the Marrian levels of analysis within the context of the interface problem, i.e. the problem of explaining the relation between the personal and different levels of subpersonal explanations. It might still be unclear how Marr's levels of analysis and the active inference framework can be used to address the interface problem. I maintain that expanding upon the previous analysis of the active inference framework through the lens of Marr's levels should dispel such worries.

The personal level, as it is standardly construed, involves thinking about the mind through the lenses of our folk-psychological abilities. At that level, we are discussing agents that possess beliefs, desires, different forms of consciousness, personality traits, and perform actions within their environments (see, e.g. Chappell, 2023; Westfall, 2022). The subpersonal levels are typically construed as, among other things, providing vertical explanations, i.e. explanations about how the personal level abilities are constituted by cognitive capacities as they are studied in psychology and cognitive (neuro)science (see, e.g. Drayson, 2012). Given this standard view, it is easy to see how to apply Marrian levels of analysis to think about the relation between the personal and the subpersonal. Personal level is that at which we detect higher-order psychological abilities that enable agents to act in their environments. This level is, thus, appropriately captured by Marr's computational analysis, which addresses questions about the function a psychological capacity performs. Given that subpersonal levels, among other things, provide vertical explanations of *how* personal level abilities are formed and constituted, this level can be analyzed in terms of algorithms and physical implementations. Of course, the feasibility of connecting personal and subpersonal levels via Marr's analysis requires the ability to computationally analyze the personal level. However, the viability of accomplishing this was demonstrated by leveraging the active inference framework. This framework posits that the mind operates by minimizing free energy and additionally provides a clear view on how the mind can be investigated from algorithmic and implementational standpoints (see also Sprevak & Smith, 2023).

A further concern might be that the active inference framework cannot provide an adequate account of personal level states and processes. Specifically, some worry that active inference cannot capture the standard roles that co-

native states, such as desires, play in explaining action and other cognitive phenomena (see, e.g. Klein, 2018; Jurjako, 2022). Indeed, in active inference, action is cast in terms of prior beliefs about policies and preferred outcomes of behaviors. Given this, the equations of active inference eschew mentioning of rewards, desires, and goals, and explain action in terms of prior beliefs (see, e.g. Hohwy, 2013, p. 89; Parr *et al.*, 2022, p. 84; Smith, Ramstead, *et al.*, 2022). Thus, the general worry is that by casting cognition in terms of free energy minimization and Bayesian beliefs, the framework lacks the resources for capturing mental phenomena that involve purely motivational or conative states as they are standardly understood in folk-psychological accounts of agency (for discussion, see Clark, 2020).

However, to think this would involve a confusion between the mathematical description of cognitive processes offered by active inference and a (folk-)psychological description of the same processes. The fact that the FEP equations do not employ terms mentioning desires, preferences or goals, does not mean that they do not refer to functional roles that at the personal level we associate with desires, preferences or goals. As compellingly shown by Ryan Smith and colleagues (2022, p. 8), once we properly distinguish the mathematical level of description from the psychological interpretations of the relevant equations, it should become clear how active inference is compatible and even illuminates some aspects of motivational personal states and processes. To show this, consider how decision-making is typically modeled within the active inference framework.

Active inference models decision-making and planning in terms of minimization of *expected* free energy, usually denoted by G . This is because variational free energy F (see above Equation 4) depends on past and present sensory observations, while to model full-blown decision-making and planning we must also consider future states and observations that actions will produce. Here, decision-making and subsequent actions are understood as processes of inferring the optimal policy that satisfy prior beliefs (e.g. Smith, Friston, *et al.*, 2022). In the context of future-oriented actions, prior beliefs will additionally encompass beliefs about the preferred states and outcomes, determined by an organism's phenotype or preferences. In this regard, expected free energy minimization will involve generative models based on which inferences about optimal policies will be performed. Accordingly, action selection is understood as an inference about which behaviors and policies will produce Bayesian beliefs about states of the world that the organism prefers to observe via its sensorium.

Similarly to variational formulations, expected free energy can also be expressed in several equivalent ways. A structurally analogous equation to (4), goes as follows (the equation is adopted from Parr *et al.*, 2022, pp. 33, 72):

$$G(\pi) = - \underbrace{E_{Q(\tilde{h}, \tilde{s}|\pi)} \left[DKL \left[Q(\tilde{h}|\tilde{s}, \pi) \parallel Q(\tilde{h}|\pi) \right] \right]}_{\text{Information gain}} - \underbrace{E_{Q(\tilde{s}|\pi)} \left[\log P(\tilde{s}|C) \right]}_{\text{Pragmatic value}} \quad (5)$$

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. $\tilde{h}$ denotes a sequence of hypothesized external states that evolve over time and produce or will produce the sequence of observations. In the first term, $E_{Q(\tilde{h}, \tilde{s}|\pi)}$ 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 $DKL[Q(\tilde{h}|\tilde{s}, \pi) \parallel Q(\tilde{h}|\pi)]$, while in the second term, $E_{Q(\tilde{s}|\pi)}$ 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 ($Q(\tilde{h}|\pi)$). Minimizing this term encourages the agent to bring its beliefs about external (hypothesized) states in line with the expected external states under the chosen policy. Importantly for our discussion is that the functional role captured by the first term can be associated with commonsensically construed belief-like states whose role is to represent the states of the environment.

In contrast, the second term $E_{Q(\tilde{s}|\pi)}[\log P(\tilde{s}|C)]$ 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 (see Parr *et al.*, 2022, pp. 73-74; Smith, Ramstead, *et al.*, 2022, p. 81). This provides the sense in which (prior) beliefs in the active inference framework, besides personal cognitive states, can also represent notions that refer to goal states, such as ends, values, desires, wants, and so on. Crucially, prior beliefs do not eliminate these motivational states; instead, they encode them through their distinct functional roles in perceptual and active inferences. Thus, despite these states being called beliefs, depending on their role in the relevant equations, they can be thought of as implementing personal level beliefs, desires or other psychological states, traits, and associated processes (for other examples and illuminating discussion, see Smith, Ramstead, *et al.*, 2022).

To further illustrate how such equations can be used to model higher-level psychological abilities, we can consider that equation (5) can be interpreted as giving a free energy formulation of the exploration and exploitation trade-off. The exploration-exploitation trade-off refers to the balance that must be struck when making every-day decisions between exploring new options and exploiting known ones to maximize rewards or values of outcomes. For example, when choosing a restaurant for dinner, exploring a new one might be a risky choice, but potentially leading to a great new culinary experience. On the other hand, sticking to familiar restaurants typically guarantees a consistently good meal.

In terms of equation (5), this aspect of decision-making can be understood as follows. The left term is often called information gain. It determines the explorative behavior, i.e. how much the organism will seek out new information to resolve uncertainty (i.e. reduce surprisal) about state variables. In our example, this involves determining how much a person would be prepared to try out a new restaurant to learn something new (i.e. what the culinary experience would be like). In this regard, the probability distributions in the left term could also be considered as encoding epistemic curiosity, as it determines the preparedness of an agent to resolve uncertainty about some situation regardless of its current pragmatic value. The right term is called pragmatic value since its probability distributions and parameters determine what kind of observations an organism expects to receive. Thus, it can be understood as encoding different types of goals, including folk-psychological constructs such as different types of conative states (see Parr *et al.*, 2022, pp. 73-74; Smith, Ramstead, *et al.*, 2022, p. 81). In general, agents can minimize their expected free energy by adopting exploratory behaviors, driven by relatively higher values of information gain thereby satisfying their epistemic needs for resolving uncertainty about some states of affairs. Alternatively, they have the option to engage in more exploitative actions, guided by increased pragmatic values that encode preferences for achieving desired outcomes.

These considerations, arguably, indicate that the active inference framework has resources to capture cognitive states, processes, and abilities as they are construed in folk-psychological explanations. This also suggests that active inference can offer a robust framework for understanding cognitive processes at the personal level and that viewing it from the prism of Marr's levels of analysis can further illuminate the relationship between personal and subpersonal levels of explanation.

7. Conclusion

In this paper, I have explored whether Marr's levels of analysis can be used as a general framework for understanding and explaining cognitive phenomena. In a discussion of this issue that has received insufficient attention, Bermúdez argues that many psychological capacities lack algorithmically implementable functions, limiting the utility of Marr's levels of analysis. Moreover, he argues that even if particular non-modular capacities could be analyzed through Marr's levels, this approach cannot be applied to the mind as a whole. To counter these arguments, I offered an example from cognitive science showing how non-modular capacities can be clarified through Marr's levels and drew on the active inference framework, which posits that the primary function of cognition is to minimize free energy. This framework enabled me to articulate the general function of the mind and consider how it can be algorithmically implemented by

physical processes. I conclude that Bermúdez has not offered principled reasons for thinking that Marr's levels of analysis cannot be used for understanding the mind as a whole.

Acknowledgments

Initial versions of this paper have been presented at the 30th SIUCC conference in Valencia (Spain) that was dedicated to José L. Bermúdez's work. I wish to thank the participants, especially José, for offering valuable comments on the early version of the paper. Thanks also to Victor Verdejo and Matheus Valente, and others for organizing such a great event. Other versions of this paper were presented at the 11th ECAP in Vienna and the TIPPS Workshop 1 in Rijeka. I thank the participants for their helpful comments. Special thanks go to Stephen Mann for reading and providing constructive comments on several versions of this paper. Finally, I would like to thank the two anonymous reviewers for *Theoria* for their critical and constructive feedback on earlier versions of this paper.

The final revisions of this paper were completed during my time as a visiting fellow at the Faculty of Philosophy, University of Cambridge. I am grateful to Professor Richard Holton for his generous hospitality and support. This paper is an outcome of the project TIPPS (<https://sway.cloud.microsoft/xtU0m67IbRR0LAiJ?ref=Link>), funded by the Croatian Science Foundation under grant HRZZ-IP-2022-10-1788. My work is also supported by the University of Rijeka under grant uniri-iskusni-human-23-14.

REFERENCES

- Andrews, K., Spaulding, S., & Westra, E. (2021). Introduction to folk psychology: pluralistic approaches. *Synthese*, 199(1-2), 1685-1700. <https://doi.org/10.1007/s11229-020-02837-3>
- Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, 36(3), 30. <https://doi.org/10.1007/s10539-021-09807-0>
- Baggio, G., van Lambalgen, M., & Hagoort, P. (2015). Logic as Marr's computational level: four case studies. *Topics in Cognitive Science*, 7(2), 287-298. <https://doi.org/10.1111/tops.12125>
- Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, 7(2), 312-322. <https://doi.org/10.1111/tops.12141>
- Bermúdez, J. L. (2000). Personal and sub-personal; a difference without a distinction. *Philosophical Explorations*, 3(1), 63-82. <https://doi.org/10.1080/13869790008520981>
- Bermúdez, J. L. (2005). *Philosophy of psychology: A contemporary introduction*. Routledge.
- Bermúdez, J. L. (2011). *Decision theory and rationality*. Oxford University Press.
- Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press. <https://doi.org/10.1017/CBO9781107279889>
- Bermúdez, J. L. (2020). *Frame it again: New tools for rational decision-making*. Cambridge University Press. <https://doi.org/10.1017/9781108131827>
- Bermúdez, J. L. (2022). *Cognitive science: An introduction to the science of the mind* (4th ed.). Cambridge University Press. <https://doi.org/10.1017/9781009064880>

- Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new Markov blankets. *Behavioral and Brain Sciences*, 45, e183. <https://doi.org/10.1017/S0140525X21002351>
- Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. *Journal of Mathematical Psychology*, 81, 55-79. <https://doi.org/10.1016/j.jmp.2017.09.004>
- Chappell, S.-G. (2023). Is consciousness gendered? *European Journal of Analytic Philosophy*, 19(1), SI8-13. <https://doi.org/10.31820/ejap.19.1.7>
- Chirimuuta, M. (2024). From analogies to levels of abstraction in cognitive neuroscience. In K. Robertson & A. Wilson (Eds.). *Levels of explanation* (pp. 200-221). Oxford University Press.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Clark, A. (2020). Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy*, 98, 1-15. <https://doi.org/10.1080/00048402.2019.1602661>
- Colombo, M., & Knauff, M. (2020). Editors' review and introduction: levels of explanation in cognitive science: from molecules to culture. *Topics in Cognitive Science*, 12(4), 1224-1240. <https://doi.org/10.1111/tops.12503>
- Colombo, M., & Wright, C. (2017). Explanatory pluralism: an unrewarding prediction error for free energy theorists. *Brain and Cognition*, 112, 3-12.
- Colombo, M., & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*. <https://doi.org/10.1007/s11229-018-01932-w>
- Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: the role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, 7(2), 243-258. <https://doi.org/10.1111/tops.12132>
- Dänzer, L. (2023). The personal/subpersonal distinction revisited: towards an explication. *Philosophy*, 98(4), 507-536. <https://doi.org/10.1017/S0031819123000220>
- Davidson, D. (2001). Mental events. In his *Essays on actions and events* (pp. 207-224). Oxford University Press.
- Dennett, D. C. (1969). *Content and consciousness*. Routledge.
- Dennett, D. C. (1981). True believers: the intentional strategy and why it works. In A. F. Heath (Ed.). *Scientific explanation: Papers based on Herbert Spencer Lectures given in the University of Oxford* (pp. 150-167). Clarendon Press.
- Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In T. K. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing* (pp. 148-160). Frankfurt am Main: MIND Group. <https://doi.org/10.15502/9783958573109>
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415-434. <https://doi.org/10.1016/j.neuron.2012.01.010>
- Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives*, 26(1), 1-18. <https://doi.org/10.1111/phpe.12014>
- Drayson, Z. (2017). Modularity and the predictive mind. In T. K. Metzinger & W. Wiese (Eds.). *Philosophy and predictive processing* (Vol. 12). MIND Group. <http://www.predictive-mind.net/DOI?isbn=9783958573130>
- Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97-115. <https://doi.org/10.1007/BF00485230>
- Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT Press.

- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, 29(1), 1-49. https://doi.org/10.1162/NECO_a_00912
- Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, 125, 88-97. <https://doi.org/10.1016/j.neubiorev.2021.02.003>
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Hornsby, J. (2000). Personal and sub-personal: a defence of Dennett's early distinction. *Philosophical Explorations*, 3(1), 6-24. <https://doi.org/10.1080/13869790008520978>
- Isomura, T., Kotani, K., Jimbo, Y., & Friston, K. (2023). Experimental validation of the free-energy principle with in vitro neural networks. *Nature Communications*, 14(1), 4547. <https://doi.org/10.1038/s41467-023-40141-z>
- Jurjako, M. (2022). Can predictive processing explain self-deception? *Synthese*, 200(4), 303. <https://doi.org/10.1007/s11229-022-03797-6>
- Jurjako, M. (2023). Uloga Marrovih razina objašnjenja u kognitivnim znanostima. *Nova Prisutnost*, XXI(2), 451-466. <https://doi.org/10.31192/np.21.2.13>
- Jurjako, M. (2024). Are mental dysfunctions autonomous from brain dysfunctions? A perspective from the personal/subpersonal distinction. *Discover Mental Health*, 4(1), 62. <https://doi.org/10.1007/s44192-024-00117-x>
- Kirchhoff, M. D., Kiverstein, J., & Robertson, I. (2022). The literalist fallacy and the free energy principle: model-building, scientific realism, and instrumentalism. *The British Journal for the Philosophy of Science*, 720861. <https://doi.org/10.1086/720861>
- Kiverstein, J., & Sims, M. (2021). Is free-energy minimisation the mark of the cognitive? *Biology & Philosophy*, 36(2), 25. <https://doi.org/10.1007/s10539-021-09788-0>
- Klein, C. (2018). What do predictive coders want? *Synthese*, 195(6), 2541-2557. <https://doi.org/10.1007/s11229-016-1250-6>
- Mallory, F. (2024). Generative linguistics and the computational level. *Croatian Journal of Philosophy*, 24(71), 195-218. <https://doi.org/10.52685/cjp.24.71.5>
- Mann, S. F., Pain, R., & Kirchhoff, M. D. (2022). Free energy: a user's guide. *Biology & Philosophy*, 37(4), 33. <https://doi.org/10.1007/s10539-022-09864-z>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- Murphy, G. L. (2019). On Fodor's first law of the nonexistence of cognitive science. *Cognitive Science*, 43(5), e12735. <https://doi.org/10.1111/cogs.12735>
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. The MIT Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- Peebles, D., & Cooper, R. P. (2015). Thirty years after Marr's vision: levels of analysis in cognitive science. *Topics in Cognitive Science*, 7(2), 187-190. <https://doi.org/10.1111/tops.12137>
- Pezzulo, G., Parr, T., & Friston, K. (2022). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1844), 20200531. <https://doi.org/10.1098/rstb.2020.0531>

- Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.
- Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, 41(9), 1017-1023. <https://doi.org/10.1068/p7299>
- Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, 22(8), 889. <https://doi.org/10.3390/e22080889>
- Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, 77(4), 477-500. <https://doi.org/10.1086/656005>
- Shea, N. (2018). *Representation in cognitive science*. Oxford University Press. <https://doi.org/10.1093/oso/9780198812883.001.0001>
- Smith, R., Friston, K. J., & Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107, 102632. <https://doi.org/10.1016/j.jmp.2021.102632>
- Smith, R., Ramstead, M. J. D., & Kiefer, A. (2022). Active inference models do not contradict folk psychology. *Synthese*, 200(2), 81. <https://doi.org/10.1007/s11229-022-03480-w>
- Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science*, tops.12704. <https://doi.org/10.1111/tops.12704>
- Verdejo, V. M., & Quesada, D. (2011). Levels of explanation vindicated. *Review of Philosophy and Psychology*, 2(1), 77-88. <https://doi.org/10.1007/s13164-010-0041-0>
- Westfall, M. (2022). Constructing persons: on the personal-subpersonal distinction. *Philosophical Psychology*, 1-30. <https://doi.org/10.1080/09515089.2022.2096431>
- Zednik, C. (2018). Mechanisms in cognitive science. In S. Glennan & McKay Illari (Eds.). *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 389-400). Routledge.

MARKO JURJAKO is an Associate Research Professor at the University of Rijeka. During the 2024/2025 academic year, he was a visiting scholar at the Faculty of Philosophy, University of Cambridge, UK. His current research primarily focuses on the philosophy of cognitive science, philosophy of psychiatry, and philosophy of law. He is the author of *Normative Reasons from a Naturalistic Point of View* (FFRI Press). His work has appeared in journals such as *Erkenntnis*, *Perspectives on Psychological Science*, *Biology & Philosophy*, *Synthese*, and the *International Journal of Psychiatry and Law*. He currently serves as Editor-in-Chief of the *European Journal of Analytic Philosophy* and President of the Croatian Society for Analytic Philosophy.

ADDRESS: University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy and Division of Cognitive Sciences, Sveučilišna avenija 4, 51000 Rijeka, Croatia. E-mail: mjurjako@ffri.uniri.hr – ORCID: <https://orcid.org/0000-0002-7252-8627>



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 / © 2026 UPV/EHU Press



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

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:

- §1 Symmetry, privacy and “I”-thoughts: A reply to Valente and Verdejo (p. 284-295)
- §2 ‘Naturalized sense data’ revisited: A reply to Brown (p. 296-301)
- §3 Elusiveness and the phenomenology of the self: A reply to Gallagher and Serrahima (p. 301-310)
- §4 Inner speech and intentional ascent: A reply to Bernués and Jorba (p. 310-316)
- §5 Preferences, frames, and reasons: A reply to Dohrn and Fisher (p. 316-329)
- §6 Active inference, the free energy principle, and Marr’s tri-level hypothesis: A reply to Jurjako (p. 330-335)

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 kinesthesia (information derived from bodily awareness about limb position and bodily movement)
- visual kinesthesia (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 $\langle a, t \rangle$ $\square$
 $\rightarrow \exists x$ [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, (**)

(**) $\exists x$ [There is an appropriately formed episode of token-reflexive thinking in context $\langle a, t \rangle$,
 $\square \rightarrow x$ 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 (*hupolēpsis*) 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 $\sim p$. 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 give 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 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-dimensional 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 of 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.

Gallagher 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 misidentification (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 composito*.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 Vygotskian 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 composito* (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 composito*).

tional ascent are themselves far less widespread than they are commonly taken to be²¹). 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 overinterprets 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 Untersuchungen* 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-cyclical preferences, 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-cyclicity. 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-cyclicity.

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-cyclicity.

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-cyclicity 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-cyclicity 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 Dohrn 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 possible by eliminating quasi-cyclicity. 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-cyclicity. 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.

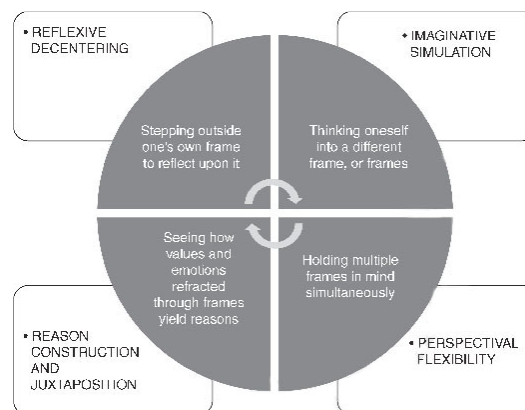


Figure 1

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 Tversky 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 Kullback-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 Mifflin.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin. 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

Theoria

AN INTERNATIONAL JOURNAL FOR THEORY,
HISTORY AND FOUNDATIONS OF SCIENCE



SUMMARY

Volume 41/2, May 2026, pp. 111-341

Articles

Valentine SIMPSON (Universitat de València), Matheus VALENTE (UNED), «The philosophy of José Luis Bermúdez: an introduction», *Theoria*, 2026, 41/2, 113-121

Victor M. VERDEJO (Pompeu Fabra University), «Thoughts about oneself to share in context: meeting Bermúdez's challenge», *Theoria*, 2026, 41/2, 122-135

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

Matheus VALENTE (UNED), «Asymmetries between ‘you’ and ‘I’», *Theoria*, 2026, 41/2, 136-156

When I tell you, ‘I’m tired’, you may report what I’ve just said as, ‘You’re tired’. Intuitively: one thought, two perspectives. Yet if a thought's identity depends on its perspective, then ‘you’ and ‘I’ couldn't possibly express the same thought. In this paper, I examine the perspectival dimensions of thought and conclude that first-person thoughts are private, apprehensible only by their owners.

Derek H. BROWN (University of Glasgow), «From immediate perception to perceptual belief», *Theoria*, 2026, 41/2, 157-178

Bermúdez (2000) offers a theory that explains referential directness via perceptual immediacy, and argues that an opposing theory cannot do the same and thus should be rejected. I reply to his criticism, and in turn argue that his theory can only explain The Problem of Perception via postulation a veil of perception, something he sought to avoid.

Shaun GALLAGHER (University of Memphis, University of Wollongong), «Why a perceptual zero-point amounts to nothing», *Theoria*, 2026, 41/2, 179-195

The concept of a “zero-point,” the bodily origin of the first-person perspective or the egocentric spatial frame of reference through which we encounter the world, is pervasive in phenomenological accounts of perception. I will argue that the zero-point is a complete abstraction that misleads the analysis of perception and action.

Carlota SERRAHIMA (University of Valencia), «Self-consciousness, ψ and ϕ », *Theoria*, 2026, 41/2, 196-214

This paper contributes to an integrative account of psychological and bodily self-consciousness (respectively, “ ψ -ownership” and “ ϕ -ownership”). It presents, first, a framework for the study of ψ - and ϕ -ownership based on José Luis Bermúdez’s. Second, it critically examines one aspect of Bermúdez’s approach to self-consciousness, namely his explanation of bodily IEM, on the grounds that it undermines the integrative project.

Daphne BERNUÉS MOSS, Marta JORBA GRAU (Universitat Pompeu Fabra), «Bermúdez’s view on inner speech: a critical assessment», *Theoria*, 2026, 41/2, 215-230

In this paper, we assess Bermúdez’s theory of inner speech as necessary for intentional ascent and argue that it is 1) empirically untenable and psychologically unrealistic, and that 2) given what it requires from inner speech, it runs into a dilemma where inner speech becomes either superfluous or inaccurately described. We finish by gesturing towards a more encompassing view of inner speech in thinking.

Daniel DOHRN (Università degli Studi di Milano), «Can quasi-cyclical preferences be rational?», *Theoria*, 2026, 41/2, 231-244

One important dimension in the work of José Luis Bermúdez concerns broadening our ideal of deliberational rationality beyond the narrow confines of standard rational decision theory. In this vein, he (2021, 2022) presents two claims: (i) there are ‘quasi-cyclical preferences’. (ii) Quasi-cyclical preferences can be rational. These claims contradict received wisdom in decision theory. I shall critically discuss both claims.

Sarah A. FISHER (Cardiff University), «Discursive deadlock and the social (media) life of frames», *Theoria*, 2026, 41/2, 245-261

Bermúdez diagnoses many cases of discursive deadlock over complex social and political issues as clashes of frames. He prescribes to agents a course of epistemic due diligence, whereby they adopt multiple alternative frames. I argue that his analysis, while valuable, is importantly incomplete. Frames also signal identities and affiliations—and that calls for structural changes to discursive environments.

Marko JURJAKO (University of Rijeka), «How general are Marr’s levels of analysis? An assessment of Bermúdez’s view», *Theoria*, 2026, 41/2, 262-282

This paper evaluates José L. Bermúdez’s claim that Marr’s levels of analysis apply only to subpersonal modular processes, not to the mind as a whole. It argues that recent advancements, particularly the active inference framework based on the free energy principle, challenge this view and demonstrate that Marr’s framework can also illuminate the relationship between personal and subpersonal explanations in cognitive science.

José LUIS BERMÚDEZ (Texas A&M University), «Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako», *Theoria*, 2026, 41/2, 283-338

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.

GUIDELINES FOR AUTHORS

THEORIA will only consider original papers in the areas of competence of our editorial board, never exceeding 10.000 words. We will try to make a decision on your paper within eight weeks.

We will only accept electronic submissions. For further Instructions, please, visit <http://www.ehu.es/theoria>

EDITORIAL PROCEDURE

THEORIA is a peer-reviewed journal. The editors, together with the competent members of the editorial board, will screen the submissions and decide whether to send them to referee. Every paper will be then assessed by two (blind) reviewers. On this basis, the editors will make the final decision. The list of reviewers is regularly published in the journal. Our archives in the editorial platform are open to research agencies (such as the Spanish FECYT) for external verification of the quality of the review process.

COPYRIGHT NOTICE

THEORIA is published under a Creative Commons Licence. Authors who publish with this journal agree to the following terms:

- Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.
- Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.
- Authors are permitted and encouraged to post their work online (e.g., in institutional repositories or on their website) prior to and during the submission process, as it can lead to productive exchanges, as well as earlier and greater citation of published work.

SUBSCRIPTION AND BACK ISSUES

THEORIA is published 3 times a year. Annual subscription rates for 2010 are: €50 for individuals and €70 for institutions. Single Issues can be ordered for €20 each. A Cumulative Index covering the numbers 21-50 is available for €3. For ordering information or any enquiry concerning subscriptions, please contact: publicaciones@ehu.eus.

Publishers

This journal is published jointly by the University of the Basque Country and the Center for Analysis, Logic, and Legal Informatics (CALIJ) in Donostia/San Sebastián, Spain.

CALIJ



THEORIA is associated with
the Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España (SLMFCE)
and the Sociedad Española de Filosofía Analítica (SEFA)

THEORIA

An International Journal for Theory, History and Foundations of Science

Vol. 41/2 • May 2026

Dep. Legal: SS - 820/85

ISSN 0495-4548

e-ISSN 2171-679X

Articles

Valentine SIMPSON, Matheus VALENTE <i>The philosophy of José Luis Bermúdez: an introduction</i> _____	113
Víctor M. VERDEJO <i>Thoughts about oneself to share in context: meeting Bermúdez's challenge</i> _____	122
Matheus VALENTE <i>Asymmetries between 'you' and 'I'</i> _____	136
DEREK H. BROWN <i>From immediate perception to perceptual belief</i> _____	157
Shaun GALLAGHER <i>Why a perceptual zero- point amounts to nothing</i> _____	179
Carlota SERRAHIMA <i>Self-consciousness, ψ and ϕ</i> _____	196
Daphne BERNUÉS MOSS, Marta JORBA GRAU <i>Bermúdez's view on inner speech: a critical assessment</i> _____	215
Daniel DOHRN <i>Can quasi-cyclical preferences be rational?</i> _____	231
Sarah A. FISHER <i>Discursive deadlock and the social (media) life of frames</i> _____	245
Marko JURJAKO <i>How general are Marr's levels of analysis? An assessment of Bermúdez's view</i> _____	262
José Luis BERMÚDEZ <i>Self, inner speech, rationality and explanation: replies to Valente, Verdejo, Brown, Gallagher, Serrahima, Bernués and Jorba, Dohrn, Fisher and Jurjako</i> _____	283
Summary _____	340