Introspection and Identification

Carl Montan

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Humanistiska teatern, Thunbergsvägen 3C, Uppsala, Thursday, 4 December 2025 at 09:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Associate Professor Léa Salje (University of Leeds).

Abstract

Montan, C. 2025. Introspection and Identification. 70 pp. Uppsala: Uppsala University. ISBN 978-91-506-3150-0.

This thesis concerns key questions in the epistemology and semantics of the mental. How are we to explain the distinctive way in which I know that I am in pain – when I know it, as we might say, 'from the inside'? How should we understand attitude attributions such as 'Sam believes that Anne is Pam'? That is, how should we understand introspection and identification?

Paper I addresses the familiar fact that each person has special access to her own mental states. In introspectively knowing that I am in pain, I know this in a way no one else can. What explains this? Drawing on a framework developed by David Enoch and Joshua Schechter, I argue that introspective de se knowledge – e.g. knowledge that I am in pain – is indispensable, whereas introspective knowledge that Carl is in pain, even if possible, would not be. This difference in indispensability explains why introspective knowledge is necessarily limited to the de se case, and thereby accounts for special access.

Paper II challenges pluralist theories – views that posit multiple capacities for introspection – which distinguish between active and passive forms of introspective knowledge. I argue that such views face serious and under-appreciated difficulties, which ultimately motivate rejecting the very idea of active introspective knowledge. I further show that these problems sharply constrain the overall space of viable pluralist theories.

Paper III examines what I call the asymmetry phenomenon: sentences like 'Sam believes that Anne is Pam' and 'Sam believes that Pam is Anne' can differ in truth-value, despite the symmetry of identity. I develop a revisionary semantic account on which ‘be’ sometimes functions to specify who someone is. I argue that this account better explains the observed asymmetries than standard views, which treat ‘be’ as either predicative or synonymous with numerical identity.

Keywords: introspection, special access, pluralism, be, Cumming's puzzle, asymmetry phenomenon

Carl Montan, Department of Philosophy, Theoreretical Philosophy, Box 627, Uppsala University, SE-75126 Uppsala, Sweden.

ISBN 978-91-506-3150-0
URN urn:nbn:se:uu:diva-570061 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-570061)

List of Papers

This thesis is based on the following papers.

I. Montan, C. Special Access to Mental States. *Manuscript.*

II. Montan, C. Introspection and Division of Labor: A Challenge to Pluralist Theories of Active Knowledge. *Manuscript.*

III. Montan, C. Asymmetry and the Meaning of ‘Be’. *Manuscript.*

Contents

Acknowledgements

There are many people I would like to thank. First and foremost, I am deeply grateful to my supervisors, Matti Eklund and Andreas Stokke. Their patience, insight, and encouragement have been invaluable throughout the process. I have learned a great deal from their philosophical acumen, precision, and remarkable eye for detail. If there are any good ideas in this dissertation, I cannot help but think that, in one form or another, I owe them to Matti and Andreas. That said, I remain solely responsible for how these ideas are presented here. I also wish to thank Åsa Wikforss for extremely helpful comments on the drafts I have sent her.

Writing a dissertation in philosophy tends to be solitary work. Still, I cannot emphasize enough how much help I have received from so many people who made this work feel far less solitary. I have had the pleasure of presenting my work several times at the Higher Seminar in Theoretical Philosophy in Uppsala. Everything contained in this dissertation has been presented there at least once – and many things that ultimately did not make it into the final version were discussed there as well. I am deeply grateful to all the participants for their insightful feedback and for helping me discern which ideas and arguments were worth keeping and which were best left aside. Additionally, I was fortunate to be part of the PhD-net, a network for doctoral students in philosophy, and to have had the opportunity to present my work in the philosophy departments of Bielefeld, Bologna, Bucharest, and Helsinki. I learned a great deal on these occasions, and the comments I received helped improve my work on many levels.

Being a PhD student at the Department of Philosophy in Uppsala has been thoroughly enjoyable. It has easily been the best job I have ever had, and I find it hard to imagine a better place to write a dissertation in philosophy. The collective breadth and depth of philosophical knowledge among the faculty, postdoctoral researchers, and fellow PhD students are inspiring in themselves; equally important are the openness, curiosity, and kindness they display. I have often arrived at – and left – work with a smile on my face. I am not entirely sure who deserves credit for this atmosphere, but it seems reasonable to assume that much of it is due to the department's leadership. I therefore wish to extend special thanks to the faculty members who have each served as

Head of Department during my time in Uppsala: Folke Tersman, Matti Eklund, and Elisabeth Schellekens Dammann.

Pär Sundström served as the external opponent at my final seminar, and I am very grateful for his incisive and constructive comments. Pär was generous with his time, encouraged open-minded discussion, and offered thoughtful criticism that I greatly enjoyed engaging with. Equally important, the dissertation in its present form has benefitted immensely from his careful feedback. The same holds for the detailed and constructive comments I received from the department's internal readers, Jessica Pepp and Pauliina Remes. Their insightful criticisms helped me recognize problems I had not fully appreciated before and inspired several revisions – sometimes through concrete suggestions – that have shaped the final presentation of this dissertation. Put bluntly, without Pär, Jessica, and Pauliina, this dissertation would have been considerably worse.

During my time as a PhD student I had the opportunity to spend the autumn of 2019 at the Department of Philosophy at the University of Helsinki. I am grateful to Maria Lasonen-Aarnio for hosting my stay there and for providing invaluable feedback on the final paper of this dissertation. I would also like to thank Studentkårens resestipendium for making my stay in Helsinki possible.

Apart from the support from Studentkårens resestipendium, my work on this dissertation has also received crucial financial assistance from the Börjessons Äldre, Henschens, and Stalins scholarships.

Besides the people already mentioned, I would like to thank the following (in alphabetical order): Tobias Alexius, Per Algander, Karl Bergman, Daniel Berntson, Sofia Bokros, Alex Byrne, Erik Carlson, Ekrem Çetinkaya, Daniel Drucker, Karl Ekendahl, Katharina Felka, Daniel Fogal, Anna Folland, Maria Forsberg, Nils Franzén, Åke Gafvelin, Susanne Gauffin, Anandi Hattiangadi, Caroline Heycock, Jaakko Hirvelä, Fabian Hundertmark, Jens Johansson, Ragnhild Jordahl, Kasper Kristensen, Sebastian Lutz, Johan Löfström, Nikhil Mahant, Guilherme Marques Pedro, Sanna Mattila, Aidan McGlynn, Sebastiano Moruzzi, Anna Nyman, Jonas Olson, Niall Paterson, Olof Petterson, Andrew Reisner, Sebastián Andres Reyes Molina, Olle Risberg, Simon Rosenqvist, Henrik Rydéhn, Jonathan Shaheen, Daniel Stoljar, Maria Svedberg, Elena Tassoni, Ulrika Valdeson, Rebecca Wallbank, Tobias Wilsch and Nick Wiltsher. I have benefited, philosophically and otherwise, from our discussions.

Finally, my deepest gratitude goes to my friends and family. To my friends – Anna, Axel, Bill, Björn, Christopher, Eric, Erik, Erika, Frida, Ian, Jessica, Joel, Jonas, Jonas, Josia, Juice, Li, Ludvig, Mani, Mike, Robin, Saam, Sandra,

Sigrid, and Shabane – thank you for your friendship and for keeping life outside philosophy rich and joyful.

I am deeply grateful to my parents, Per and Ia, for helping me with virtually everything, both big and small. I also want to thank my wife, Tytti, for her unwavering support and for all the laughter we have shared. Even during times when I have struggled with my work, she has always made me laugh – often at myself, as she is particularly skilled at mocking me.

Last of all, I want to thank my children, Laban and Helmi. They have, in their own way, contributed greatly to this dissertation – mostly by preventing me from actually working on it. Sometimes by literally closing my laptop when I have tried to write. Thank you both for doing an excellent job of keeping my priorities straight.

Introduction – Kappa

§1. Introduction

As the term 'self-knowledge' is used in the contemporary philosophical literature, it concerns a person's propositional knowledge of her own mental states. Specifically, it is propositional knowledge where the content is thought of under *a first-person mode of presentation*: knowledge that *I* am in a mental state.

The significance of self-knowledge as a philosophical topic largely arises from the intuitive idea that we each know our own mental states in a way that others cannot. Following standard usage, I refer to such knowledge as 'introspective', and I will use the term 'special access' as follows:

> SPECIAL ACCESS For all persons *S* and all mental states *M* such that *S* is in *M*, *S* has special access to *S*'s being in *M* iff there is a way of knowing *W* such that *S* can know that *S* is in *M* in way *W* and for all persons $S^* \neq S$ it is metaphysically impossible that S^* knows that *S* is in *M* in way *W*.

In simpler terms, introspection provides individuals with a way of knowing their own mental states – a way that necessarily applies only in their own cases.[1]

This dissertation consists of the present introduction ('Kappa') and three papers that explore different aspects of the epistemology and the semantics of the mental. Papers I and II (see below) explore different aspects of introspective knowledge of mental states, while Paper III (see below) examines how some of these mental states are reported in natural language.

I. 'Special Access to Mental States'

It is widely held that we have special access to our own mental states. However, there has been little systematic discussion of what explains it – that is, what explains the fact that you cannot know, for example, that I am in pain in

[1] For endorsement, see e.g. Alston (1971), Bar-On & Wright (2024: 2), Byrne (2005; 2018: 4-9), Davidson (1987: 441-442), Farkas (2008: 17-18), Fernández (2013: 5), Gertler (2011: 65-67), Shoemaker (1996: 25).

the same way as I do when I know it through introspection? In this first paper, I present and defend an answer to this question. I begin with the observation that much of our introspective knowledge is *basic*, where *S*'s knowledge that *p* is basic iff *S* knows that *p* and the fact that *S* knows that *p* is independent of anything else *S* knows. Given that special access involves basic knowledge, I propose that we look for a general theory of basic knowledge – a theory that explains how basic knowledge of all different varieties is possible – to account for our special access. For this purpose, I adopt a theory of David Enoch and Joshua Schechter (2006; 2008) on the justification of basic belief-forming methods.

The central idea of Enoch and Schechter is that a method of belief-formation yields basic justification when it is indispensable for successful engagement in a rationally required project – such as understanding the world, deliberating about action, or evaluating one's own reasoning. On this view, the epistemic status of a belief-forming method derives entirely from its indispensability for such projects. Because basic knowledge is possible only through methods that make such success possible, the indispensability account provides a principled and uniform explanation of basic knowledge.

I show that this framework yields a natural explanation of special access. Introspective *de se* knowledge, for example knowledge that *I am in pain*, is indispensable for rationally required projects such as rational action. By contrast, introspective knowledge that *Carl is in pain* – even if it were possible – would not be indispensable in this way. This difference explains why only the former can constitute basic introspective knowledge. In short, special access is a consequence of the rational role that basic introspective *de se* knowledge plays in our lives.

I argue that this provides the best available explanation of special access, comparing it to a range of alternative accounts. To test these other accounts, I introduce the possibility of *shared mental states*.[2] Consider a hypothetical case where conjoined twins, Tom and Arnold, have partially overlapping brains, resulting in some of their mental states being shared. For instance, their pains and tickles have both Tom and Arnold as bearers. I use the case of Tom and Arnold to challenge many otherwise seemingly natural explanations of special access. I show that they struggle to explain why Tom cannot introspectively know that Arnold is in pain. By contrast, the indispensability account explains this by saying that there is no rationally required project for which introspective knowledge that Arnold is in pain would be indispensable; therefore, Tom

[2] For discussion of shared mental states, see Armstrong in Armstrong & Malcolm (1984: 110f), Kang (2022), Roelofs (2019; 2024) and Sutton (2014).

cannot introspectively know that Arnold is in pain. If there is a better explanation of special access, it remains to be found.

II. 'Introspection and Division of Labor: A Challenge to Pluralist Theories of Active Knowledge'

Recent debates on self-knowledge have seen a shift from *monism* – the view that a single capacity underwrites all introspective knowledge – toward *pluralism* – the view that there are multiple introspective capacities. Pluralist theories vary in their details and motivations.[3] The present paper examines in detail the pluralist accounts of Matthew Boyle (2009; 2011) and Annalisa Coliva (2016; 2017; 2019). Their proposals are motivated by two ideas. First, the central claim that some kinds of mental states – in particular attitudes such as desires and beliefs – are known actively, through deliberation or reasoning. Second, that other kinds of mental states – such as sensations – are not known actively but passively. To account for the full range of our introspective knowledge, they posit two distinct capacities, one for active and one for passive knowledge.

To assess these theories, I introduce a key distinction between *strict* and *non-strict* divisions of labor between introspective capacities. A strict division of labor allows no overlap in which capacity can target a given mental state, while a non-strict division permits such overlap. This distinction sets up a dilemma for Boyle and Coliva.

If they adopt a strict division of labor, their accounts are incomplete: some introspective knowledge goes unexplained. Attempts to repair this incompleteness either add *ad hoc* complexity or erode the rationale for positing multiple capacities. Under a strict division of labor, their accounts are therefore unsatisfactory.

If they adopt a non-strict division of labor, they face the challenge of explaining the absence of *contradictory introspective judgments*. By *contradictory introspective judgments* I mean two (roughly) simultaneous introspective judgments, where one is of the form 'I am F' and the other 'I am not F'. The problem for views adopting a non-strict division of labor is, first, that the most natural explanation of this absence is unavailable once two overlapping capacities are posited, and second, that remaining explanatory strategies either lack independent philosophical motivation or require denying the fallibility of introspection. Under a non-strict division of labor, their accounts are again unsatisfactory.

[3] For different pluralist proposals, see e.g. Butler (2013), Carruthers (2010; 2011) and Wright in Bar-On & Wright (2024).

Since neither horn is palatable, I argue that these results motivate rejecting the very idea of active introspective knowledge.

Finally, I consider the wider implications for pluralism about introspection. I argue that while no single problem I raise for Boyle and Coliva is a problem for pluralism as such, the combined force of the problems significantly narrows the range of viable pluralist theories.

The two papers just summarized address our introspective knowledge of mental states. The third paper shifts focus, examining how some of these mental states are reported in natural language. A quick link between the two topics can be forged via the phenomenon of *immunity to error through misidentification*. As Sydney Shoemaker (1968) famously observes, when I introspectively judge that I am in pain, it is impossible that my judgment be in error for the following reason: that I have mistaken someone else's pain for my own. In §5 below I discuss how this might be explained, and what those explanations might show about introspection. But here and now I want to focus on something narrower. When we consider the relevant attributions of mistake – confusing one thing with another, misidentifying some person as someone else – an intriguing asymmetry emerges. There is a difference between misidentifying someone else as oneself and misidentifying oneself as someone else. Paper III explores the nature of this asymmetry and proposes an account of the phenomenon.

III. 'Asymmetry and The Meaning of 'Be''

The *asymmetry phenomenon*, as I refer to it, can be illustrated by a pair of sentences originally discussed by Sam Cumming (2008):[4]

(1) Biron thinks Katherine is Rosaline.
(2) Biron thinks Rosaline is Katherine.

Although identity is symmetric, (1) and (2) can differ in truth-value. This result is surprising if, as tradition holds, 'be' flanked by referential terms always expresses identity.

A further datum sharpens the point. When 'is' in (1) and (2) is replaced with 'is identical to', the asymmetry appears to vanish:

(1=) Biron thinks Katherine is identical to Rosaline.

[4] The asymmetry phenomenon is not restricted to belief. It also shows up with other attitude verbs. Moreover, related phenomena are found with counterfactuals, questions and so-called identificational sentences (an example is: 'that is Rosa'). For further discussion, see §6 below.

(2=) Biron thinks Rosaline is identical to Katherine (from Kocurek 2018).

This indicates that 'be' in (1) and (2) cannot simply mean *be identical to*.

To account for the asymmetry phenomenon, I propose that the order of referential terms in the relevant clauses makes a semantic contribution. Since numerical identity is inherently symmetric, (1) and (2) require another semantic value for 'be'. I propose a new denotation for 'be' that builds the order of referential terms directly into its meaning. It does so by encoding claims about who someone is, rather than asserting numerical identity:

WH-IDENTITY [[be]] = λyλx. who x is is y.

WH-IDENTITY takes an ordered pair of expressions 'y' and 'x' of type <e>, and yields the true iff who x is is y. WH-IDENTITY allows the order of terms around 'be' to make a semantic contribution, and so accounts for the readings on which (1) and (2) have different truth-values. More concretely, on my suggestion (1) is seen as an answer to the question: who does Biron think that Katherine is? In contrast, (2) is seen as an answer to the question: who does Biron think that Rosaline is? Different questions, different answers, different truth conditions: the meaning of 'be' as WH-IDENTITY explains the asymmetry phenomenon. I further show how this accounts for other data that belong to the asymmetry phenomenon.

My proposal naturally invites a question: if the relation encoded in WH-IDENTITY and numerical identity are distinct relations, what distinguishes them? I have little definitive to say. In Paper III I argue that the linguistic evidence strongly supports treating them as distinct. However, a fully informative account of their difference remains to be worked out.

In conclusion, this paper argues that the traditional view of 'be' as always expressing numerical identity (when referential terms flank 'be') fails to account for certain linguistic data. Even if important details remain to be worked out, the WH-IDENTITY proposal emerges as the best explanation of the asymmetry phenomenon.

Having outlined the three papers of this dissertation, I now explain how the remainder of this introduction proceeds. §2 describes how recent theorizing about self-knowledge has developed in critical response to what it has perceived as an inaccurate picture of the mind and our knowledge of it – one often described, with a good deal of simplification, as the philosophical inheritance from Descartes. §3 presents what is often regarded as the key explananda for a theory of introspective knowledge. §4 evaluates the main theories of introspection. §5 examines immunity to error through misidentification (IEM) as

it relates to introspection. Finally, §6 turns to issues in the philosophy of language, focusing on attitude ascriptions and the meaning of 'be'.

These sections connect with the arguments presented in the papers as follows. §§2–4 provide an overview of the discussion of self-knowledge, forming the background to Papers I and II. Paper I revolves around Hume's claim (discussed briefly in §2 and again in §5) and evaluates whether current theories of introspection (covered in §4) can explain special access. While Paper I offers a concise statement of these theories, §4 supplies further detail for readers seeking additional background.

Similarly, Paper II draws on theories of introspection discussed in §4. §4.1 and §4.5 contain discussion of Shoemaker's argument against the possibility of self-blindness, which provides context for Paper II's response to an anti-pluralist argument. Moreover, some of the challenges I raise for Boyle's and Coliva's pluralism draw support from widely accepted views about introspection presented in §3.

§5 links the themes of Papers I, II, and III by examining why introspective first-person judgments (judgments of the form 'I am F') are immune to error through misidentification. This bears directly on Paper I's question of how to explain special access – in particular, how to explain introspective *de se* knowledge if Hume's claim is correct. Finally, §6 offers background for the asymmetry phenomenon discussed in Paper III.

§2. A Simplified Picture and Its Contemporary Rejection

In this section I sketch a widely invoked simplified picture of the mind and of self-knowledge that contemporary philosophers take themselves to have inherited from earlier thinkers. This picture – more a construction than a piece of historical scholarship – continues to stand as a foil for current theorizing. I mention some of the canonical figures and some of the core ideas that – rightly or wrongly – have come to be linked together.

In *Our Knowledge of the Internal World* (2008), Robert Stalnaker gives a clear statement of the picture that contemporary theorizing has been shaped in response to:

> The Cartesian internalist begins with the contents of his mind – with what he finds by introspecting and reflecting. That is what is unproblematic; these are the things and the facts we know directly. The internalist's problem then is, how do we move beyond these to form a conception of an external world, and how are we able to know that the world beyond us answers to the conceptions that we form (2008: 2-3).

On the picture that Stalnaker labels the *Cartesian internalist* outlook, the central epistemological challenge is to explain how knowledge of the external world is possible.[5] By contrast, knowledge of the internal world is treated as unproblematic. On this picture, introspection is taken to provide direct and highly reliable access to one's own mental states, and functions as the natural starting point for further theorizing rather than as a phenomenon that itself requires explanation.

On the simplified picture, René Descartes stands as the originator of this outlook. He is taken to have confronted global skepticism with the powers of introspection by first finding certainty in knowledge of his own existence as a thinking thing: "I am thinking, therefore I exist" (1985a [1637]: 32/127).[6] After securing certainty about his own existence, Descartes is then understood as having extended the foundation of certain propositions to the rest of his own mind.[7] The simplified picture associates Descartes with three claims about introspection:[8]

> INDUBITABILITY Necessarily, for all persons *S* and all propositions *p*, if *S* introspectively believes that *p*, it is impossible for *S* to doubt that *p*.
>
> INFALLIBILITY Necessarily, for all persons *S* and all mental states *M*, if *S* introspectively believes that *S* is in *M, S* is in *M.*
>
> OMNISCIENCE Necessarily, for all persons *S* and all mental states *M*, if *S* is in *M*, *S* introspectively knows that *S* is in *M.*

As the relevant flavor of modality is metaphysical, INDUBITABILITY says that it is metaphysically impossible to doubt what one introspectively believes; INFALLIBILITY that it is metaphysically impossible to have false introspective beliefs about one's own mental states and OMNISCIENCE that it is

[5] For a similar discussion, which emphasizes the problem of other minds, see Bilgrami (2006: 3-15).

[6] Since I am – and most readers will be – accustomed to the translation 'I think therefore I am' is worth noting that the authoritative Cottingham-Stoofhoff-Murdoch renders the phrase as 'I am thinking, therefore I exist'.

[7] Variations on this picture recur in Byrne (2018: 17) and Shoemaker (1996: 52-53). For relevant discussion in Descartes' own work, see Descartes (1985b [1641]: 29/19).

[8] For representative statements of this received picture, see Bar-On & Wright's careful attribution: these claims are "associated (perhaps erroneously) with Descartes [namely, that] that each one's mental life constitutes a totally transparent inner theatre, with an audience, necessarily, of one" (2024: 3). Less carefully, Boghossian says: "Many extravagant claims have been made about our capacity to know our own minds. Descartes, who was responsible for the worst excesses, taught many subsequent generations that self-knowledge was both infallible and exhaustive" (2008 [1989]: 140).

metaphysically impossible to fail to introspectively know one's own mental states. INDUBITABILITY plays no further part in this dissertation; the other two claims are scrutinized in §3.

Subsequent philosophers are likewise folded into this picture. John Locke is often taken to uphold the Cartesian conviction that introspection is unproblematic. He describes introspection as an *inner sense*: "the perception of the operations of our own minds within us" (1975 [1690] II: 105). David Hume is commonly read as following Locke in treating introspection as a kind of perception. On this picture, Hume – the great skeptic of the era – could doubt almost everything (God, causes, substances) but introspection. Or, rather, he *almost* makes an exception for introspection: for while he does not doubt what introspection presents, he observes that it does not present him with a self (this is what I referred to as *Hume's claim* in §1):

> For my part, when I enter most intimately into what I call *myself*, I always stumble on some particular impression or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never can catch *myself* at any time without a perception, and never can observe anything but the perception. [---] If any one upon serious and unprejudiced reflection, thinks he has a different notion of *himself*, I must confess I can no longer reason with him (Hume 2000 [1739-1740]: 164).

On a natural understanding, Hume takes this observation to show that the self does not exist (or that the concept *I* is empty).[9] While this conclusion has always been controversial, the underlying observation that there is no phenomenology of the self analogous to the phenomenology of pain is relatively uncontroversial and has shaped much theorizing about introspection (for further discussion, see §5 below and Paper I).[10]

Moving to the 20th century, Bertrand Russell is widely understood as having given introspection a crucial role in his epistemology. In *The Problems of Philosophy* (1912), he contrasts the dubiousness of external-world knowledge with the indubitability of introspection:

> My knowledge of the table as a physical object […] is not direct knowledge. Such as it is, it is obtained through acquaintance with the sense-data that make

[9] I am unsure if Hume ultimately takes a stand on the matter whether I can introspectively know that I am in pain or whether knowledge is limited to there being pain (or something like that). Howell (2023: ch. 3) distinguishes between different philosophically interesting interpretations.
[10] For instance, Immanuel Kant accepted Hume's observation that introspection (what Kant calls "inner sense") does not reveal "the soul itself as an object" (1998 [1781/1787]: A22/B37) but argues that this "I or he or it (the thing) which thinks" (A346/B404) nevertheless exists. Kant infers its existence on the basis of two claims, (i) that we have knowledge of the empirical world, and (ii) that the self is a necessary condition for all knowledge.

> up the appearance of the table. We have seen that it is possible, without absurdity, to doubt whether there is a table at all, whereas it is not possible to doubt the sense-data (1912: 47)

On the simplified picture, Russell's epistemological program retained strong influence on philosophers for a few decades. Eventually, the theoretical landscape underwent change. Gilbert Ryle (1949) attacked the picture of the mind as a hidden inner theater open to a single spectator. Ludwig Wittgenstein (1953) attacked the idea of a private language. These critiques marked a turning point. The general outlook in philosophy of mind and epistemology shifted in opposition to what was widely regarded as a problematic legacy.

As Stalnaker emphasizes, contemporary philosophy no longer begins from a conception of the certainty of the inner. Instead, it begins with "the world we find ourselves in, and with what either common sense or our best scientific theories tell us about it" (2008: 3). From this vantage point, it seeks "to explain how our objective conception of the world can be a conception of a world that contains things like us who are able to think and experience it in the way that we do" (3). Our own minds are no longer regarded as the obvious, unproblematic starting points of all inquiry. Skepticism about strong claims like INFALLIBILITY and OMNISCIENCE is now widespread, and many theorists reject the idea that introspection is a form of inner perception. The focus has shifted from the claims associated with Descartes toward a more detailed critical examination of what introspection is and how it operates.

In the next section, §3, I introduce what most contemporary theorists regard as the key phenomena, the explananda, for a theory of introspective knowledge, setting the stage for the evaluation of these theories in §4.

§3. Explananda

A common claim in the literature on self-knowledge is that we have privileged first-person access to our own mental states. Alex Byrne (2005) helpfully distinguishes two key claims that have not always been clearly separated in earlier discussions.

The first concerns what he calls *peculiar access*: "One's knowledge of one's mental states is *peculiar* in comparison to one's knowledge of others' minds" (Byrne 2005: 81).

> One has a special method or way of knowing that one believes that the cat is indoors, that one sees the cat, that one intends to put the cat out, and so on, which one cannot use to discover that someone *else* is in the same mental state. Our access to others' minds is importantly similar to our access to the non-psychological aspects of our environment: one can come to know that the cat

> is indoors by seeing that it is, and one can likewise come to know that the cat wants to be fed and that Fred wants sushi deluxe. Our peculiar access to our own minds is not like this: one can come to know that one wants sushi deluxe without observing oneself at all (Byrne 2005: 81).

In the terminology I introduced above, this is the claim that we have *special access* to our own mental states.[11] This claim is typically not argued for.[12] It is usually taken as a datum – something to be accepted unless compelling reasons to reject it emerge. In Paper I, I examine whether the possibility of *shared mental states* constitutes such a reason. I argue that it does not.

The second claim that Byrne identifies concerns the reliability of introspection. Beliefs about one's own mental states "acquired through the usual route", i.e. introspection, "are more likely to amount to knowledge" than other persons' beliefs about those mental states (Byrne 2005: 80). For example, my belief that Fred wants sushi deluxe based on testimony or observation is generally less secure than Fred's introspective belief that he wants sushi deluxe. Byrne calls this *privileged access* (Byrne 2005: 80), a term I adopt as well.[13]

Byrne emphasizes that special and privileged access can come apart in both directions. First, having a special method of knowing that *p* does not entail that my belief that *p* is more likely to amount to knowledge than someone else's belief that *p* formed by other means.[14] For instance, if I use my own spy app installed in Fred's phone to track Fred's whereabouts, I know where Fred is in a way no one else can (imagine that only I have the spy app). However, since the app is unreliable (I am an average app constructor), Fred's own beliefs about his whereabouts are more likely to amount to knowledge than mine. Second, the fact that my belief that *p* may be more likely to amount to knowledge than your belief that *p* does not entail that I have special access to *p*. For instance, as an expert on wildflowers in Scandinavia, my beliefs about them are more likely to amount to knowledge than Fred's, but our method of forming belief – using vision – is the same.

---

[11] Note that Byrne takes no stand on the modal flavor of 'impossible'. Authors such as Armstrong (1993 [1968]: 124), Kang (2022: 4) and Lycan (1996: 48-49) takes the impossibility to be something like nomological impossibility. Going forward, I assume the orthodox position that it is *metaphysically impossibility* that is relevant.

[12] An exception is Farkas (2008: 17-19), although what she presents as an argument could (I think) just as well be described as an illustration. She invites the reader to focus on a present mental state and imagine what it would be like to know that some distinct person was in a corresponding mental state. She concludes: "If you are anything like me, then it will be clear that you cannot learn about their mental states in the way you did about your own" (2008: 19).

[13] It is also endorsed by e.g. Bar-On & Wright (2024: 2), Fernández (2013: 6) and Gertler (2011: 61-65).

[14] The individuation of belief-forming methods is a theoretical problem that I set aside for the sake of simplicity. For a statement of the "Generality problem", see Conee & Feldman (2004: 145-166). For a proposed solution, see Lyons (2019).

One reason to distinguish sharply between special and privileged access is that even when access is special, it may vary in how privileged it is. For instance, I have special access both to the fact that I am in pain and the fact that I see a goat. My former belief is more likely to amount to knowledge than the latter. Still, the way I know that I see a goat, 'from the inside', is different from how you know that I see a goat (by observing me, my eyes, and the goat).

Byrne's concept *privileged access* is quite modest. In §2, I introduced the much stronger claim INFALLIBILITY, associated with earlier theorizing about introspection:

> INFALLIBILITY Necessarily, for all persons *S* and all mental states *M*, if *S* introspectively believes that *S* is in *M, S* is in *M.*

While earlier theorists may have endorsed INFALLIBILITY, few contemporary philosophers do. Instead, they typically endorse something like the claim that we have privileged access. The following counterexamples to INFALLIBILITY, falling into five groups, make clear why.

1. *Emotions, feelings and sensations.* Some emotions, feelings and sensations are hard to get right.[15] While my introspective belief that I am sad has a high likelihood of being true, my introspective belief that my current sadness is a symptom of depression could be mistaken. A newly amputated person may wrongly think that her foot hurts. While she is correct that she is in pain, she is wrong to think that her foot is the source of the pain.

2. *Priming cases*. Christopher Hill (1988) recounts a famous example of a priming case originally introduced by Rogers Albritton:

> A college student […] is being initiated into a fraternity. He is shown a razor, and is then blindfolded and told that the razor will be drawn across his throat. When he feels a sensation he cries out: he believes for a split second that he is in pain. However, after contemplating the sensation for a moment, he comes to feel that it is actually an experience of some other kind. It is, he decides, a sensation of cold. And this belief is confirmed when, a bit later, the blindfold is removed and he is shown that his throat is in contact with an icicle rather than a razor (1988: 18).

To show how priming can lead to false introspective belief, we must first make explicit that the student's initial introspective belief that he is in pain is false – the sensation is one of cold, not pain. That clarified, what the case shows is that we can be tricked into forming false introspective belief.

3. *Borderline cases*. Pains hurt and itches itch. Imagine a continuum where extreme pain is at one end of the spectrum, and extreme itch is at the other.

[15] For this claim, see e.g. Burge ([1996] in 2013: 77 n. 8).

Somewhere along this continuum there is a range of unclear cases where I am neither itchy nor in pain. In such borderline cases, two introspective errors seem possible. Either (a) I introspectively believe that I am in pain, or (b) I introspectively believe that I am itchy; both beliefs can be false.

4. *Detailed cases.* Gilbert Ryle raised the 'Problem of the Speckled Hen' against A.J. Ayer. Roderick Chisholm gives the problem its classical formulation:

> Let us consider the visual sense-datum which is yielded by a single glance at a speckled hen. The datum may be said to "comprise" many speckles. [---] Our problem pertains to the question: how many speckles does the datum comprise? C.I. Lewis has written that there can never be "positive bafflement in the presence of the immediate, because there is here no question which fails to find an answer" (Lewis 1929: 128). But what of the question about the number of speckles? If we judge that there are forty-eight, it would seem, at first consideration at least, that we might very well be mistaken [;] we are unable to make a reliable judgment about what we *do* notice (1942: 368).

Chisholm's point as he states it concerns sense-data. But there is a more general point independent of such theoretical posits. Imagine that you see a speckled hen (a *real* hen) and the side facing you has forty-eight speckles. You introspectively judge that you see forty-one speckles, and your judgment is false.[16]

5. *Factive mental states*. Introspection can lead to false beliefs about so-called factive mental states. I may introspectively believe that I *see* a church (but I am wrong, because what I see is a barn), that I *remember* having read Strawson's *Individuals* (but I am wrong, because I have only intended to read it) or that I *know* that Oslo is the biggest capital in Scandinavia (I have gotten the geographical facts wrong).[17]

As noted above, these well-known counterexamples explain why INFALLIBILITY has few defenders today. How reliable introspection ultimately is remains an open question; what seems hard to deny is that introspection, in general, is more reliable than perception or testimony when it comes to knowledge of mental states. Since this is all that is required for privileged access, and since the reliability of introspection plays only a limited role in this

[16] Cautiously retracting from 'seeing' to (the philosophically sophisticated) 'having a visual impression as of' is not guaranteed to save you from error; if there is too much detail you may be wrong about how things seem to you.

[17] Some philosophers resist the idea that *knowing is a mental state*, the view famously defended by Williamson (2000: 22-48, where opposition to the idea is also discussed). I wish to remain neutral on whether knowing is a mental state. Still, it is worth noting that if one denies that knowing is a mental state, then – given the natural claim that we can introspectively know that we know – introspective knowledge would not always be knowledge of mental states. There is a trade-off here, which raises questions I cannot enter here.

dissertation, I proceed under the assumption that Byrne's modest claim about the reliability of introspection is unproblematic.

While most theorists agree that a theory of introspective knowledge should explain why we have *special* and *privileged access* to our own mental states, many contend that more needs to be explained. There is a persistent idea that, at least for some restricted range of mental states, we tend to know that we are in them whenever we are.[18] This natural conviction plausibly explains why some philosophers have been drawn to the stronger claim OMNISCIENCE, repeated below:

> OMNISCIENCE Necessarily, for all persons *S* and all mental states *M*, if *S* is in *M*, *S* introspectively knows that *S* is in *M*.

However, OMNISCIENCE is too strong, as several well-known counterexamples show. Examining them is instructive since it helps point to a weaker more plausible principle in the vicinity of OMNISCIENCE.

1. *Regress.* OMNISCIENCE implies that each piece of introspective knowledge is accompanied by a distinct piece of introspective knowledge, generating a vicious regress. (While I will later consider potential revisions of OMNISCIENCE that avoid this problem, for now let us continue to gather counterexamples to get more material for that revision.)

2. *Poverty and lack of sophistication.* Animals and young children have mental states but lack the requisite concepts to have introspective knowledge. They can suffer pain, but without concepts they cannot know that they are in pain.

3. *The unconscious and inaccessible mental items.* While Freud's theory of unconscious mental states is controversial, the existence of unconscious mental states and processes is widely accepted. If Descartes had an unconscious desire to kill his father, he did not introspectively know that he did.[19]

4. *Distractions*. Introspective knowledge can be blocked by distraction. Shoemaker notes:

> Examples abound of cases in which, allegedly, a person is in pain without being aware of the pain [or introspectively knowing that she is in pain]: the wounded soldier who in the heat of the battle is oblivious to his pain, the injured athlete who does not notice his pain until he is put on the bench, and so on (1996: 227).

[18] As Bar-On & Wright put it: "we tend, in the round, to know what there is to know: our own mental attributes […] do not, in general, elude our awareness, although those of others may often do so" (2024: 2).

[19] Modular processes are similarly inaccessible.

In fact, the same point holds even for minor inattention.[20]

5. *Anti-luminosity.* Williamson argues that there are no non-trivial luminous conditions (2000: 126-129):

> Consider a morning on which one feels freezing cold at dawn, very slowly warms up, and feels hot by noon. One changes from feeling cold to not feeling cold, and from being in a position to know that one feels cold to not being in a position to know that one feels cold. If the condition that one feels cold is luminous, these changes are exactly simultaneous. Suppose that one's feelings of heat and cold change so slowly during this process that one is not aware of any change in them over one millisecond. Suppose also that throughout the process one thoroughly considers how cold or hot one feels. One's confidence that one feels cold gradually decreases. One's initial answers to 'Do you feel cold?' are firmly positive [---] one's final answers are firmly negative (Williamson 2000: 96-97).

Somewhere along the line, after firm assertion but before firm negation, Williamson says: "one gives neutral answers such as 'It's hard to say'" (97). The conclusion is that one is not always in a position to introspectively know what mental state one is in.[21]

In light of these counterexamples, OMNISCIENCE cannot be sustained. However, they suggest directions for refinement. One might restrict the principle to occurrent mental states, thereby excluding unconscious and inaccessible mental states. One might also restrict it to creatures with requisite concepts. Annalisa Coliva proposes one such revision:

> SELF-INTIMATION Given C-conditions (including concept possession, cognitive well-functioning, alertness and attentiveness and to the exclusion of unconscious and purely dispositional mental states), if one has a given mental state *M*, one will be in a position to judge or believe (or both) that one has it (Coliva 2016: 60).[22]

Coliva's addition of 'C-conditions' does important work. Adding them eliminates the problems for OMNISCIENCE listed under 2 above (the cognitively unsophisticated), the problems listed under 3 (Freudian unconscious and inaccessible mental states are excluded) and those listed under 4 (distractions). And because it appeals to being *in a position to judge or believe*, rather than knowing, it circumvents the problem listed under 1 (regress).

---

[20] For further examples and discussion, see Byrne (2018: 37-38).

[21] The literature on Williamson's anti-luminosity argument is vast. For criticism, see e.g. Berker (2008), Conee (2005), Weatherson (2004) and Wright in Bar-On & Wright (2024: 39-42). For defenses and discussion, see Manley (2007) and Srinivasan (2013).

[22] Coliva calls the principle 'Strong Transparency'. Since the term 'transparency' is used for other claims and theses, I have changed it to 'self-intimation' (which is often used in this way).

Nevertheless, SELF-INTIMATION is not without difficulties. First, for psychological reasons, it is questionable whether one is always in a position to make a (true) judgment about the mental states one is in – even if the C-conditions are met.[23] Second, since beliefs themselves are mental states (*M* in SELF-INTIMATION), the principle seems to imply that I am also in a position to form introspective belief about having this very (second-order) belief, and so on. Yet, at some point my capacity for higher-order belief gives out, and as formulated, SELF-INTIMATION does not accommodate this.

As it stands, SELF-INTIMATION needs further refinement. Still, something in its vicinity is widely taken to be correct. Theories of introspection are commonly assessed, at least in part, by their ability to account for a suitably qualified principle of this sort. In §4, I follow this practice.

One final feature often highlighted in discussions of introspective knowledge is its *directness*. Directness is typically characterized negatively, as knowledge that is not acquired by way of some additional step. For instance, Donald Davidson makes this point when he emphasizes that introspective knowledge is knowledge acquired "without appeal to evidence" (1987: 441). The term 'evidence' is used in different ways in epistemology. Here the relevant use is this: *e* is evidence for the truth of *p* if *e* is an indication of the truth of *p.* When Davidson sees that the streets are wet and infers that it is raining, the wet streets are his evidence.[24] If his belief that it is raining amounts to knowledge, it is *evidential* knowledge. By contrast, when Davidson knows that he *believes* that it is raining, there is no proposition that *p* that serves as evidence for his belief. In this sense, his introspective knowledge is non-evidential.

Crispin Wright makes a complementary point. He observes that when I utter "I am tired" (sincerely and with understanding):

> [t]he demand that somebody produces reasons or corroborating evidence for such a claim about themselves – 'How can you tell?' – is always inappropriate. There is nothing they might reasonably be expected to say (Wright 2001: 321).

The point is this. Ordinarily, when we ask someone to justify a belief, we expect them to provide a *reason* that supports it. For example, if asked why she believes that it is raining, a person can offer the reason that the streets are wet,

[23] For example, consider standing beliefs such as my belief that Vienna is the capital of Austria. Even when alert, attentive, and conceptually competent, I do not seem to always be in a position to make the judgment that I have this belief. This casts doubt on whether Coliva's C-conditions suffice for always being in a position to make true judgments about one's own mental states.

[24] Some may prefer to say that it is my seeing (or my having a visual experience as of) the streets being wet that is my evidence. It does not affect the point whether evidence is mental (states of seeing or believing) or non-mental (states of the world, like a street being wet).

thereby justifying her belief. By contrast, in the introspective case no such reason seems available. If we ask the same person how she knows *that she believes* that it is raining, the natural reply is simply: "I just (know that I) believe that it is raining". This does not provide a reason; it merely expresses the belief.

From these discussions, a preliminary result suggests itself: *S*'s knowledge that *p* is direct only if it is not based on evidence or reasons. This formulation states only necessary conditions on direct knowledge; it deliberately sidesteps the attempt to capture every aspect of directness. However, it suffices to fix ideas for what follows.

Understood in this way, the directness of introspective knowledge shares some properties with certain kinds of perceptual knowledge.[25] Some perceptual knowledge – for example, that the streets are wet – can also be direct, in that it does not rely on evidence or reasons.

This is a good place to make a comment on the terminology I will adopt in this introduction. Some theorists speak of positive epistemic support that is distinct from evidence and reasons and label it *warrant*.[26] I will use that term in this introduction. In this sense, introspective beliefs and some perceptual beliefs are warranted. Evidence will play no further role in my discussion. However, reasons will. I use 'reason' in the following way: *q* is a reason for *S* to believe that *p* iff (i) *S* knows, believes, or is aware that *q,* (ii) *q* is distinct from *p*, and (iii) *q* supports *p*.

To summarize, discussions of introspective knowledge typically focus on four key features: (1) introspection provides a special way of knowing one's own mental states (*special access*); (2) introspective beliefs are more secure than others' beliefs about one's own mental states (*privileged access*); (3) for some kinds of mental states, if certain conditions are met, one is typically in a position to introspectively believe that one is in that mental state (*self-intimation*); (4) introspective knowledge is direct. In the next section, I turn to theories of introspection and consider how they explain (or fail to explain) these features.

§4. What is Introspection?

In this section, I evaluate the main theories of introspection. I first discuss observational models of introspection, including the inner-sense theory (§4.1), and the acquaintance theory (§4.2). I proceed with non-observational theories:

[25] As Wright (2015: 70-72; 2024: 194-196) observes.
[26] See e.g. Burge [1993] & [1996] in (2013).

consciousness-based theories (§4.3), transparency theories (§4.4) and constitutivist theories (§4.5). The final summary (§4.6) brings these discussions together and categorizes the main challenges these theories face in addressing the key explananda outlined in §3.

§4.1. The Inner-sense Theory

The first family of views to consider is the inner-sense theory, which holds that introspection is a kind of perception. As Nico Silins aptly puts it, the idea is "historically common, and perhaps perennially tempting" (2013: 291). Locke famously describes introspection (his term for it was 'reflection') as "the perception of the operations of our own minds within us" (1975 II: 105):

> This source of ideas every man has wholly in himself; and though it be not sense, as having nothing to do with external objects, yet it is very like it, and might properly enough be called *internal Sense*. But as I call the other *Sensation*, so I call this *Reflection*: the ideas it affords being such only as the mind gets by reflecting on its own operations within itself (1975 II: 105).

The core claim of the inner-sense theory is that just as perception yields knowledge of the external world, an inner sense yields knowledge of one's own mental states. In contemporary philosophy, David Armstrong is a leading defender.[27] He writes:

> It is clear that we obtain such knowledge as we have of the current state of the physical world by means of sense-perception. [---] [The inner-sense theory] asserts that for the case of a person's own mind, though not for any other mind, a perception-like process goes on which yields the person information about the current state of his mind (1984: 108-109).

On Armstrong's view, mental states *cause* introspective beliefs and these beliefs amount to knowledge because the processes that generate them are reliable.

As outlined in §3, discussions of introspection tend to focus on four central explananda. I will evaluate the inner-sense theory in light of these. Overall, the inner-sense theory has some immediate intuitive appeal: it promises to explain self-intimation and directness in a straightforward way. But when we evaluate it systematically against the four explananda, the shortcomings become apparent.

Let us begin with *special access*. At first sight, the inner-sense theory seems to accommodate special access. If introspection is a perception-like faculty proprietary to the subject, it is confined to one's own mental states by

[27] For other defenders, see Churchland (2013) and Lycan (1996).

(metaphorically speaking) being locked inside the skull (cf. Armstrong (1993 [1968]: 124; 1984: 109)). This would explain why we never gain introspective knowledge of other persons' mental states.

However, special access requires more than this. It is not enough that, as a matter of fact, I do not introspectively know others' mental states; what is needed is that it be metaphysically impossible that I introspectively know that e.g. you are in pain. Since it is conceivable that the inner eye could be reconfigured or relocated to perceive someone else's mental states, the inner-sense theory fails to secure the modal force of special access.[28]

Next, initially, the inner-sense theory appears promising for explaining *privileged access*. If introspection is a kind of perception, one might expect introspection to share the reliability of ordinary perceptual faculties. Just as vision reliably delivers information about the external world, so inner vision might reliably deliver information about the mind.

However, by itself this is not enough to explain privileged access. To do so, the theory must show not only that introspection is reliable, but that it is *more* secure – more likely to generate knowledge of mental states – than ordinary perception or testimony. The inner-sense theorist's claim that inner perception is reliable, while plausible, does not by itself explain this greater security. What is needed is an account of why inner perception should be especially reliable – and that is precisely what the inner-sense model has not supplied (for discussion, see Byrne 2018: 48-49).[29]

The inner-sense theory has its strongest intuitive appeal in connection with *self-intimation* of phenomenal states. When I am in pain, I am typically in a position to introspectively believe (or know) that I am. This is a straightforward instance of self-intimation. At first sight, the inner-sense model seems to provide a natural explanation: I believe that I am in pain by perceiving the pain through an inner faculty.

The difficulty is that the range of self-intimating states is not confined to the phenomenal. I also typically know that I have a standing desire to visit China. But such states are not occurrent phenomenal events, and so the model that applies in the case of pain does not straightforwardly extend to them.

---

[28] Armstrong (1984: 110f) himself and Lycan (1996: 48-49) recognize this but apparently do not see it as a defect of the theory. Most theorists do, see e.g. Moran (2001: 33).

[29] One line of response, noted by Byrne, appeals to the relative simplicity of the causal chain involved in introspection. The transition from my first-order mental state to the introspective belief that I am in it is much shorter than the transition from another person's mental state to my belief that they are in that mental state. The idea is: shorter causal chains involve fewer opportunities for error, thereby yielding greater reliability. Yet as Byrne points out, "everything turns on the unknown details" of this hypothesis (2018: 49). While this suggestion identifies a possible structural advantage of inner perception, it still falls short of a substantive explanation of privileged access.

The inner-sense theorist might insist that the same perceptual story applies across the board – that I perceive my desire just as I perceive my pain. Yet this response strains plausibility. In the case of pain, there is a phenomenology that the inner sense might be said to detect; in the case of a standing desire, there is none. It is unclear how both could be explained by the same perceptual model.

Last, at first glance, the inner-sense theory seems well suited to account for directness. Just as I directly see a duck on a truck, I can introspectively know that I am undergoing a visual experience as of a duck on a truck by directly perceiving it. In both cases, perception appears to put me in direct contact with the objects known.

However, important disanalogies make this suggestion implausible. First, each perceptual modality – vision, audition, touch, taste and smell – has its own characteristic phenomenology. If introspection were perceptual, we would expect it to have a distinctive phenomenology. Yet, as Fred Dretske observes:

> introspection has no phenomenology or, if it [has one] it always has the *same* phenomenology as the experience one is introspecting. If one is asked to say what one's current visual experiences are like, for example, one seems able to know this without having any identifiable experiences other than the visual ones one is able to describe (1995: 62).[30]

To make the point vivid, consider the case of pain. If introspection were inner perception, we should expect two experiences whenever I know that I am in pain: the pain itself, and an introspective experience of that pain. Yet the latter does not seem to exist – casting doubt on the perceptual model.

Second, perceptual experiences characteristically allow for misidentification. I may see a man across the street, George, and misidentify him as Greg. But in introspection there is no analogous possibility of misidentification (I return to this in §5). As Shoemaker explains:

> I become aware that George has shaved his beard by seeing George, seeing that he is beardless, and identifying the man I see as George, the man I know to have been previously bearded. So the provision of identification information is an important part of the role played by awareness of objects giving us awareness of facts. But in introspective self-knowledge there is no room for an identification of oneself, and no need for information on which to base such an identification [---] This comes out in the fact that there is no possibility here of

[30] For similar claims, see Burge ([1998] 2013: 384), Moran (2001: 13-14), Shoemaker (1996: 207) and Wright in Bar-On & Wright (2024: 275-276 & 293).

> a *mis*identification; if I have my usual access to my hunger, there is no room for the thought "Someone is hungry all right, but is it me?" (1996: 210-211).[31]

In sum, Shoemaker's and Dretske's observations highlight decisive disanalogies between ordinary perception and introspection. They strongly suggest that the directness of introspection cannot be understood in the same way as the directness of perception.[32]

What we have gathered so far is enough to discard the inner-sense theory. However, before closing it is worth considering a different and oft-cited concern: that the theory misrepresents the relationship between introspection and our own minds. The most influential argument here is Shoemaker's argument from the impossibility of self-blindness. This argument also motivates Shoemaker's positive view (§4.5), and it plays an important role in Paper II, where I (on behalf of pluralism) respond to an argument against pluralism. It therefore deserves attention here.

Shoemaker (2009: 37) defines *self-blindness* as a condition in which a creature with normal intelligence, rationality and conceptual capacities – let *C* represent this cluster of properties – lacks a capacity for introspective knowledge and can only access her own mental states through third-person means.[33] Shoemaker's core claim is that if the inner-sense theory is correct, then self-blindness must be metaphysically possible (1996: 224-245). Let *P* abbreviate 'perceptual capacity' as construed by the inner-sense theory. If *P* is anything like ordinary perceptual capacities, then *C* and *P* are independent. This independence entails that a creature could possess *C* while lacking *P*, similar to ordinary blindness, or having *P* while lacking *C*, as in certain cognitive impairments. Therefore, on the inner-sense theory, it is metaphysically possible for a creature to have *C* yet lack *P* – that is, to be self-blind.

Shoemaker's next step is to deny this possibility. While Shoemaker offers several complex arguments to demonstrate this, it is the intuitive idea behind them that has gained widespread support.[34] Suppose that you are in pain and you have *C.* Pain is a feeling, and given that you have the relevant concepts,

---

[31] For objections to Shoemaker's claims about the absence of misidentification, see Macdonald (1999), and Shoemaker (1999) for a response.

[32] A related claim is worth noting. It is sometimes said that no appearance/reality distinction can be drawn for phenomenal states, see e.g. Gertler (2011: 95-96), Horgan (2012: 406), Kripke (1981: 152-154) and Wright (2001: 304). If correct, this invites at least two explanations. One is that experiences of mental states are exceptional in being immune to misrepresentation. Another, more plausible explanation, is that there are no experiences of mental states over and above the phenomenal states themselves. The latter casts further doubt on the inner-sense theory.

[33] These are more or less the same conditions that Coliva invokes (see my discussion in §3; specifically, the discussion of self-intimation).

[34] See e.g. Coliva (2016: 167-174), Dretske (2003: 11), Gertler (2011: 155-156), Parrott (2017), Stoljar (2018: 391-392). For dissent, see Lycan (1996: 16-19).

nothing seems to stand in the way of your forming the introspective judgment that you are in pain. Intuitively, in having *C*, it seems that you have everything needed to be able to introspectively judge that you are in pain. Conversely, it is very hard to see how you could lack the capacity for introspective judgment given that you feel pain and you have *C*. If this intuition is right, there is no possible world in which you are in pain and you have *C*, yet you lack a capacity for introspective belief.

Since the possibility of self-blindness is held to be deeply counterintuitive, Shoemaker's argument poses a significant challenge for the inner-sense theory.

To summarize: the inner-sense theory provides no satisfactory explanations of special access, privileged access, self-intimation or directness and it faces the additional problem that it renders self-blindness possible. In the next section, I turn to a nearby view: the acquaintance theory.

§4.2. The Acquaintance Theory

Bertrand Russell introduces the epistemic relation on which his epistemology rests, *acquaintance*, in terms of *direct awareness*: "We shall say that we have *acquaintance* with anything of which we are directly aware, without the intermediary of any process of inference or any knowledge of truths" (1912: 46). Russell sometimes describes this awareness as perceptual, claiming that one "sees" mental items such as sense-data (1912: 50). This suggests that Russell's acquaintance theory is simply a version of the inner-sense theory discussed in §4.1. If that is correct, the same difficulties apply: the theory would struggle with explaining special access, privileged access, self-intimation, and directness.[35]

However, not all acquaintance theorists embrace this perceptual gloss. For instance, Brie Gertler (2012; 2018) explicitly distances her view from inner-sense models. Her theory takes acquaintance to be a *sui generis*, non-causal, and direct relation between mental states and introspective beliefs.[36] On this view, acquaintance provides *observational awareness* of one's own mental states, and this awareness is what warrants our introspective beliefs.

We can distinguish between two versions of Gertler's theory, depending on how it construes the relationship between awareness of pain and pain itself.

---

[35] Some present-day acquaintance theorists, like Duncan (2015; 2019), are vulnerable to these objections.

[36] This contrasts sharply with Armstrong's materialist, reliabilist, and causal inner-sense theory. It is not obvious that these differences track the distinction between acquaintance and inner-sense theories as such; one could, in principle, imagine a reliabilist or even materialist acquaintance theory. Still, since Gertler explicitly formulates her view in opposition to Armstrong's, I present it in the terms she adopts.

First, suppose that the *awareness is distinct from the pain*. On this version, I have introspective knowledge of pain through being acquainted with it, where (a) acquaintance is direct awareness of the pain, and (b) this awareness is distinct from the pain itself. The ontology thus involves three items: (i) the pain, (ii) the awareness of the pain, and (iii) the introspective belief that I am in pain. This picture invites familiar worries. Is there really such an awareness over and above the pain? Why does this awareness not allow for misidentification, as perceptual awareness does? And why should it be metaphysically impossible to lack the capacity for this sort of awareness?

Second, suppose instead that the *awareness is identical to the pain*. On this version, I have introspective knowledge of pain through being acquainted with the pain, where the acquaintance with pain simply is being in pain. Although this simplifies the theory, it raises uncomfortable questions. What explanatory work does acquaintance now do? If being acquainted with pain simply is being in pain, the appeal to acquaintance risks redundancy. Indeed, the resulting view looks close to the theory we will examine in §4.3, which makes no reference to acquaintance at all. If that theory can do the same explanatory work, it should be preferred for its simplicity.

Ultimately, neither Russell's nor Gertler's versions of the acquaintance theory provide significant advancement over the inner-sense theory. If acquaintance is taken as distinct from mental states, the theory faces many of the same difficulties as the inner-sense view. If, on the other hand, acquaintance is identified with the states themselves, the notion risks doing no explanatory work.

These misgivings suggest a different strategy: perhaps we should give up the idea that introspection is a form of observation altogether. Instead, we might seek an account that treats the intimate link between conscious states and introspective belief as a different kind of relation. The next theory we consider develops this approach.

§4.3. Consciousness Based Theories of Introspective Knowledge

As said, one thing we want a theory of introspective knowledge to explain is that when I am in a phenomenally conscious state, it is typically evident to me that I am. A natural suggestion is that phenomenal states themselves somehow ground introspective knowledge – not some perception of them. Christopher Peacocke (1999) develops this idea as follows:

> Conscious thoughts and occurrent attitudes, like other conscious mental events, can give the thinker reasons for action and judgment. They do so also in the special case in which they give the thinker a reason for self-ascribing an attitude to the content which occurs to the thinker, provided our thinker is conceptually equipped to make the self-ascription (1999: 214).

Peacocke extends this account beyond paradigm reason-giving mental states, such as belief, to include states like pains, tickles, and other feelings. He writes that an "experience of pain can in ordinary circumstances be a thinker's reason for judging that he is in pain, a judgment which in those circumstances amounts to knowledge" (1999: 215).[37] I take Peacocke, together with more recent defenders (Howell 2023; Neta 2011; Silins 2013; Smithies 2019), to be committed to a principle of the following form:

> CONSCIOUSNESS BASED For all subjects *S* and all phenomenally conscious mental properties P of *S*, if *and because S* instantiates P*, S* has a reason to introspectively believe that *S* instantiates P.

Three quick comments are in order.

First, the emphasis on reasons is meant to be congenial to epistemologically internalist sympathies.[38] I say more about this below.

Second, while CONSCIOUSNESS BASED is formulated in terms of 'reasons', the account is intended to explain introspective *knowledge*.[39]

Third, there is a concern that CONSCIOUSNESS BASED needs to be paired with a substantive account of introspection – one that posits a mechanism that detects mental states.[40] Without such a detection mechanism, it is hard to see how the theory could explain knowledge at all. Most advocates of CONSCIOUSNESS BASED will resist this suggestion. Their project is precisely to explain introspective knowledge without positing a detective mechanism.[41] Still, it is natural to suspect that this points to a potential explanatory gap in the account. My critical remarks in what follows do not depend on this worry – but it is worth noting.

Evaluated against the explananda in §3, CONSCIOUSNESS BASED offers explanations of some phenomena but not others.

First, the account gives the outlines of an explanation of privileged access. The idea would be that introspective knowledge is privileged because the very states that provide introspective reasons are the states introspective beliefs are

---

[37] "Perceptual experiences and sensations, on the one hand, and so-called 'occurrent' conscious propositional attitudes, on the other, differ in many respects. But there is one property they share. They both contribute to what, subjectively, it is like to have them for the person who enjoys them" (1999: 205-206).

[38] Peacocke's suspicions about reliabilism is a recurrent theme in his book (see e.g. 1999: 240-241).

[39] This is particularly clear in Smithies (2019: 154-155).

[40] O'Brien (2012: 176) presses this worry in connection with Peacocke's account of knowledge of mental actions (e.g. judgments). See Peacocke (2012: 184) for his response.

[41] See in particular, Peacocke (1999: 215) and Smithies (2019: 154-159). Unlike Peacocke and Smithies, Neta emphasizes that (his version is of) CONSCIOUSNESS BASED is consistent with most accounts of what introspection consists in.

about. Since states and reasons coincide in this way, introspective beliefs formed on these bases cannot be false. In this respect, they contrast sharply with third-person beliefs about my mental states, which are fallible.[42]

Second, with a few additional assumptions, Peacocke's framework initially seems to offer a natural explanation of special access. As Declan Smithies develops it, the explanation is that:

> our introspective reasons for beliefs about mental states have their source in those very mental states. When I feel cold, this gives me an introspective reason to believe that I feel cold, and I'm thereby in a position to know that I feel cold by forming a belief on the basis of this reason. Other people are sometimes in a position to know that I feel cold on the basis of reasons provided by perception, testimony or inference, but they are never in a position to know that I feel cold on the basis of reasons provided by the fact that I feel cold. This way of knowing about our own mental states is different from our ways of knowing about the mental states of other people (2019: 157).

This is an appealing start. However, the account fails to capture the full modal force of special access. It is not enough to note that others cannot, in ordinary circumstances, know that I am in pain on the basis of my pain; what must be explained is that it is metaphysically impossible for others to do so. Here the theory falters as I discuss in Paper I. Recall from §1 that when I am in pain, I can introspectively know that I am in pain but not that Carl is in pain. On CONSCIOUSNESS BASED, the warrant is provided by the pain state itself, and it presumably warrants equally (see Stoljar 2012: 393): so, CONSCIOUSNESS BASED does not explain why I cannot introspectively know that Carl is in pain. The case of shared mental states makes the inability to explain special access even clearer. If I share my pain with my Siamese twin Arnold, then – given that pain is supposed to be reason-providing – the account offers no explanation for why I cannot introspectively know that Arnold is in pain. CONSCIOUSNESS BASED fails to secure special access.[43]

Next, consider directness. At first sight, CONSCIOUSNESS BASED seems to contradict it. In §3, I said that *S*'s knowledge that *p* is *direct* only if it is not

[42] However, to explain privileged access, two more claims need to be secured: first, it must actually be the case that we form our introspective beliefs in the way Peacocke describes. Second, the account must show why false introspective beliefs are in fact rare; unless the account establishes that introspective error is less frequent than is the case in third-person knowledge of mental states, it is incomplete. As stated, CONSCIOUSNESS BASED does not speak to this matter.

[43] It might be thought that CONSCIOUSNESS BASED can be reformulated to say that pain gives reason to believe that I am in pain, but not that Carl is in pain. However, such a reformulation seems to be in tension with the simple idea that animates the theory: that being in a phenomenal state suffices for introspective warrant.

based on reasons. Since CONSCIOUSNESS BASED claims that introspective knowledge is based on reasons, it seems to contradict directness.

However, this conclusion is too quick. On the characterization adopted in §3, a proposition *q* is a *reason* for believing *p* only if (i) *q* is a proposition that I know or have justification to believe, (ii) *q* is distinct from *p*, and (iii) *q* supports *p*. Peacocke and Smithies use 'reason' differently: for them, my being in pain is itself a reason for believing that I am in pain. But being in pain does not satisfy condition (i), and arguably not (ii) either.[44]

Given this terminological divergence, CONSCIOUSNESS BASED may ultimately accommodate directness. However, the conception of reasons raises a different worry. Internalist-leaning epistemologists typically operate with the characterization of reasons I have offered. By deviating from that conception, Peacocke and Smithies invite the question of whether their view genuinely preserves internalist commitments, as they claim.[45]

Moving on, last, to self-intimation, CONSCIOUSNESS BASED is explicitly framed around phenomenal states. Yet, as noted earlier, self-intimation is not confined to such states: standing states such as desires and beliefs fall within its scope as well. Smithies proposes an extension to cover belief:

> Necessarily, if you believe that *p* then you thereby have an introspective reason that puts you in a position to know by introspection that you believe that *p* (2019: 171).

At first glance this looks promising, but it raises a worry about the theory's underlying motivation. If non-phenomenal states like belief can provide introspective reasons, why place such explanatory weight on phenomenology? What distinctive work does phenomenal consciousness do if warrant arises even in its absence?

Peacocke offers a related strategy (1999: 240-242) that is vulnerable to criticism of the same kind. He suggests that when I know my beliefs introspectively, CONSCIOUSNESS BASED can explain this by appealing to a counterfactual: I would have known the belief if I had based my introspective belief about it on a phenomenally conscious mental state (a judgment). But this move is unconvincing. Explaining actual knowledge in terms of knowledge I *would* have had by a different route is rarely satisfactory. Knowing that there is butter in the fridge because I remember putting it there is not explained by the fact that I would have known it had I looked inside. Likewise, appealing to what I

[44] Perhaps condition (iii) is met, on certain interpretations: perhaps there is a sense in which the proposition *that I am in pain* supports belief *that I am in pain*. I leave that open.
[45] For this criticism, see Coliva (2008); for defence of Peacocke, see McHugh (2012).

would have known by way of a conscious judgment does not illuminate how I actually know my beliefs introspectively.

In summary, CONSCIOUSNESS BASED theories mark an advance over perceptual theories, but they leave crucial issues unresolved. In particular, they fail to secure the modal force of special access and struggle to extend self-intimation beyond phenomenal states. For these reasons, the account remains unsatisfactory as it stands. These shortcomings naturally invite alternative approaches that explain introspective knowledge without appealing to perception or to consciousness as a special source of warrant. One such approach is the theory first sketched by Gareth Evans (1982). It shifts the focus from inward awareness of mental states to outward-directed engagement with the world.

§4.4. Transparency Theories

Gareth Evans (1982) introduced what has come to be called the *transparency theory.* He writes:

> [I]n making a self-ascription of belief, one's eyes are so to speak, or occasionally literally, directed outward – upon the world. If someone asks me "Do you think there is going to be a third world war?", I must attend, in answering him, to precisely the same outward phenomenon as I would attend to if I were answering the question "Will there be a third world war?" (Evans 1982: 225).

Since Evans, the transparency idea has been developed in different ways by theorists such as Byrne (2018), Richard Moran (2001) and Daniel Stoljar (2019) – among others.[46] Evans's own remarks are somewhat programmatic and underdeveloped, so in what follows I focus less on Evans and more on subsequent developments of his ideas. As with previous theories, I assess the different transparency accounts by asking how well they accommodate the four explananda from §3. I focus primarily on how transparency theories seek to account for introspective knowledge of belief, though I briefly consider possible extensions at the end of this section.

First consider privileged access. Evans himself thought that his envisaged theory would explain it. In describing the procedure of answering the question, 'Do I believe that *p*?' by considering whether p, he writes:

> There is no question of my applying a procedure for determining beliefs *to something*, and hence no question of my possibly applying the procedure to the wrong thing. If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states: even the most determined sceptic cannot find here a gap in which to insert his knife (1982: 225).[47]

[46] Cf. Fernández (2013) and Gallois (1996).
[47] Parentheses in the original removed for readability.

While the passage is somewhat cryptic, it is clear that Evans regarded transparency as guaranteeing knowledge of one's own beliefs. But although Evans was confident that the transparency procedure guarantees knowledge of one's own beliefs, his remarks leave open a pressing question: how can looking outward at the world yield knowledge of belief at all? Transparency theorists owe us an account of why the outward-directed procedure generates knowledge rather than mere guesswork.

Byrne offers the most detailed contemporary attempt to answer this challenge, and his account promises to explain privileged access. According to Byrne, introspective knowledge that I believe that *p* is achieved through epistemic rule-following. In the case of knowledge of belief, the relevant rule is BEL, which states: "If *p,* believe that you believe that *p*" (2018: 102). According to Byrne, I follow BEL on an occasion when I infer that I believe that *p* from my recognition of the fact that p.[48] By following BEL, Byrne argues, one achieves *safe* belief,[49] and hence introspective *knowledge*. Moreover, since I cannot follow BEL without believing that *p*, following BEL is (more or less) self-verifying. And this is how Byrne's transparency theory proposes to explain privileged access: introspection is more secure than third-person methods because if one follows BEL one (almost) always ends up with a true belief.[50]

As can be seen, Byrne's account is cast within an epistemologically externalist framework. But the transparency idea is not tied to externalism. Internalism-friendly developments are possible.

Stoljar (2019) offers one. On his development of Evans's idea, outward-directed attention at bottom amounts to this. Taking p to be the case is to believe that *p*, and in considering whether p – and taking it to be the case – one's belief that *p* becomes (access-)conscious.[51] Stoljar holds that when a belief (or more generally, a mental state) is conscious one is rationally required to form the corresponding introspective belief, so long as certain conditions obtain. He captures this requirement in the following principle:

> IP It is rationally necessary that if one is in a conscious state *C*, then (given certain conditions) one will believe that one is in *C*.[52]

[48] As stipulated, one follows BEL only if one's belief that *p* is true. Problem: one can have introspective knowledge of false beliefs too. Byrne's solution: I *try to follow* the BEL-rule on an occasion, if I believe that I believe that *p* as a consequence of my false belief that *p*.
[49] Belief that could not easily have been false. For details, see Williamson (2000: 123-128).
[50] For another transparency account of privileged access, see Fernández (2013: 53-59).
[51] Stoljar (2019: 2070-2073). A mental state is *access conscious* if it is poised or ready for use in controlled processing (Block 1995: 231).
[52] IP is a simplification. Stoljar's principle says that if one is in a conscious state *C and* it matters to one whether one is in *C*, *and* one has the concepts required to believe one is in *C*, *and* one

On Stoljar's proposal, IP is one among the rational principles that govern the minds of rational agents. It belongs with more familiar principles such as the requirement that if one believes *p* and that if *p* then *q*, one will believe *q* – principles that are, in principle, accessible to reflection.

Although Stoljar does not frame IP as an account of introspective *knowledge*, such an extension is natural. If belief-formation in accordance with rational principles open to reflection is paradigmatically rational, then such belief is plausibly also knowledge-conducive, at least in the normal case. On this understanding, compliance with IP constitutes a route to knowledge.

However, even under this development, it is unclear whether IP explains privileged access. The account shows why beliefs formed in compliance with IP are rational, but not why they should be more secure than other ways of forming beliefs about mental states. Perhaps such an explanation can be developed within the framework, but in its present form it has not been supplied.

So much for privileged access. I now turn to special access. At first sight, the transparency idea looks well-suited to explain it. If I can know that I believe that *p* by attending to whether p, it is natural to think that this procedure works only for me. By attending to whether p, I can arrive at the *belief* that you believe that *p*, but it seems quite obvious that this belief cannot amount to *knowledge* that *you* believe that *p*. My attending to whether *p* provides no warrant to think anything about your mental states – it is no better than guesswork. In this rough form, then, the transparency idea promises a straightforward explanation of special access.

However, matters become more complicated once we examine particular developments of the transparency idea. To help show this, again consider shared mental states. Suppose that I share the belief that *p* with my Siamese twin Arnold, and I form the belief that Arnold believes that *p* by attending to whether p (as I do in my own case). Here Byrne's rule-following account appears to undercut rather than explain special access: my belief that I believe that *p*, and my belief that Arnold believes that *p*, formed by the same method (attending to whether p), are arguably equally safe. If one amounts to knowledge, the other should as well. In contrast, Stoljar's account at first sight looks better positioned. It seems plausible enough that only the subject who is in a conscious state is rationally required to self-ascribe it. If compliance with IP can generate knowledge, this appears to explain why only I can introspectively know that I believe that *p.* Yet closer inspection reveals a gap. To secure special access, the account must establish more than IP itself states: it must explain why it is metaphysically impossible for anyone else – such as the twin

considers whether one is in *C*, *and* … (further conditions could be added), then one believes that one is in *C* (2018: 401; 2019: 2070).

with whom I share mental states – to be governed by the same rational requirement with respect to our conscious states. Perhaps such an explanation can be developed, but as yet none has been provided.

Next, consider directness. Both Byrne and Stoljar accommodate directness. On their accounts, attending to whether p and thereby arriving at the belief that I believe *p* is not a matter of basing the belief on reasons. My consideration of whether p is neither evidence for, nor a reason to believe, that I believe *p* (see Byrne 2018: 104–106).

However, directness is arguably not accommodated by all versions of the transparency theory. Moran's (2001; 2003) proposal is a case in point. Moran holds that the transparency procedure delivers introspective knowledge because it is tied to my reasons for believing that *p*. On his view, introspective knowledge that I believe that *p* is achieved by attending to the reasons I have for believing that *p*. Apparently, then, introspective knowledge is not direct on Moran's view.

Now, this may look like another case where 'reason' is used in a special way, prompting the suspicion that Moran's view may accommodate directness after all. However, on closer inspection, I think that this suspicion is misguided. Consider introspective knowledge of belief as portrayed by Moran. I introspectively know that I believe that it rains by considering my reasons for believing that it rains. I have just seen the dog come in soaking wet, and this is my reason to believe that it rains. In that case, my seeing the wet dog is my reason to introspectively believe that I believe that it rains. Since this appears to be Moran's view, it fails to accommodate directness.[53]

Finally, we arrive at self-intimation. With respect to belief, the transparency theory looks promising. Simplified: the explanation of why I am typically in a position to introspectively know that I believe *p* is that all I need to do is to consider whether p is the case. Since this procedure is almost always available to me, the direct accessibility of my beliefs is accommodated.

However, a challenge of scope arises. Can the transparency explanation extend beyond belief to other kinds of mental states – pain, thirst, tickles, and the like?[54] If not, the transparency theory cannot provide a complete account special access, privileged access, directness and self-intimation.

On first encounter, the prospects for extension might seem uncertain. Some theorists, such as Boyle (2009), Fernández (2013), and Moran (2001), concede the limitation. They restrict their transparency theories to introspective knowledge of belief (and some other propositional attitudes), leaving the

[53] Additionally, it does not seem to describe how I introspectively know that I believe that it rains.

[54] Bar-On (2004: 114-116; cf. Bar-On & Wright 2024: 60-61) raises this critical question for Evans's account.

explanation of introspective knowledge of non-attitudinal states to other approaches. The attractions and limitations of Evans's original remarks have served as a source for pluralist views of introspection, discussed in Paper II.

Others, most notably Byrne, aim to extend the transparency framework across the board: all introspective knowledge – of every kind of mental state – is transparent.[55] This move raises a pressing question (see e.g. Bar-On 2004: 115): if introspection always works by 'looking out', what exactly am I attending to when I come to know that I am in pain, that I am thirsty, that I feel ticklish or that I want butter? There is no obvious answer. Ordinary concrete objects – cars and butter – are clearly of the wrong kind. And familiar *mental* objects do not seem worldly enough to ensure uniformity with Evans's outward-looking idea.

For the sake of space, I focus on the problem as it applies to introspective knowledge of pain. Byrne suggests an account which draws an analogy with perceptual modalities. Each perceptual modality (audition, sight, touch, gustation and olfaction) reveals the world in its own distinctive way (2018: 128-146). For instance, sight reveals 'a world of vision': a small portion of facts of the visible world. Byrne encourages us to think of pain in a similar way. What he calls "the world of pain" consists of "the totality of *p-facts*, concerning qualities of painful disturbances in the bodies of animals" (2018: 149). A *p-fact* is a "true *p-proposition*, which we can write as '$[\ldots x \ldots]_p$', where *x* is a disturbance" (149).

According to Byrne, *p-propositions* are "in one way perfectly familiar, although giving a theoretical characterization of [them] is difficult" (149):

> When I feel a pain in my foot, the curtain is raised on a segment of the world of pain: a pulsating unpleasant occurrence in my foot. The fact that $[\ldots x \ldots]_p$ specifies that segment, the scene before my nociceptors [i.e. pain detectors]. If [you] and I strike up a conversation about our gout, p-facts are a significant part of our topic (Byrne 2018: 149).

Byrne proposes that just as BEL underwrites introspective knowledge of belief, the rule PAIN underwrites introspective knowledge that I am in pain:

> PAIN If $[\ldots x \ldots]_p$, believe that you feel a pain (Byrne 2018: 149).

I follow PAIN on an occasion if I believe that I feel pain because I recognize $[\ldots x \ldots]_p$. Doing so tends to yield knowledge because the belief is safe.

[55] Evans himself (1982: 230-231) indicates that his account would explain knowledge that I am in pain too.

How satisfactory is this explanation?[56] I think that there is a genuine worry that Byrne's "worlds of pain" and p-propositions do not supply the outward-directed objects that a uniform transparency theory requires. Saying that I recognize the fact $[\ldots x \ldots]_p$, when I feel pain is clearly (at least superficially) different from saying that I recognize the fact that it hurts. And the label 'world of pain' of course gives the impression of worldliness. Yet, at a substantive level, the phenomenon still appears sufficiently mental to disrupt any hoped-for uniformity with the outward-looking procedure – however it is ultimately to be understood. The worry, in short, is that replacing "I feel pain" with "[…x…]p" looks more like relabeling than explanation.

While these concerns specifically target Byrne's generalizing move, similar worries arise for any attempt to generalize the transparency idea. Whether the framework can truly be generalized to cover all forms of introspective knowledge remains an open question.

Let us take stock. I have suggested that some versions of the transparency theory can plausibly explain special access and privileged access, and that some can also accommodate directness and self-intimation. Yet the scope problem remains pressing: it is far from clear that transparency can be extended beyond belief to cover the full range of mental states. For this reason, no version of the theory is entirely satisfactory as it stands. This invites a different kind of approach – one that does not rely on outward-directed or inward-directed procedures at all but instead locates the source of introspective knowledge in the very nature of the mentality of rational creatures. Constitutivist theories, which are the last we will look at, take precisely this route.

§4.5. Constitutivist Theories

Constitutivist theories propose that there is a tight metaphysical connection between being in a mental state and believing that one is in that mental state.[57] The idea is that, under certain conditions, the existence of a mental state itself guarantees its accessibility. Key proponents of constitutive theories include Akeel Bilgrami (2006), Coliva (2016), Shoemaker (1996; 2009) and Wright (2001; 2024).

There are many ways of articulating this constitutive connection. Some are crude and quickly run into trouble; others are more programmatic and require careful evaluation. My strategy in this section will be to work through these formulations in turn: starting with some crude proposals, continuing with

[56] For critical discussion, see Keeling (2019: 641) and Bar-On & Johnson (2025).

[57] For reasons that will emerge, it is appropriate to state this first – rough stab – thesis in terms of 'belief' instead of 'introspective knowledge', since some constitutivists deny that the relevant beliefs amount to knowledge.

more programmatic ones whose plausibility is harder to assess, and arriving finally at a version of constitutivism, close to Shoemaker's, that seems more attractive.

Starting with the straightforward claims, one sometimes finds proposals of the following sort:

> (i) It is constitutive of *certain kinds of mental states* (e.g. beliefs, sensations) that their occurrences are introspectively knowable.[58]
> (ii) It is constitutive of *intentional agency* that the mental states on which intentional actions are based are introspectively knowable.[59]

Formulations (i) and (ii) are underspecified in two important respects: first, they do not say which kinds of mental states are covered; second, they do not say whether introspective knowledge is supposed to be constant (that one always introspectively knows), typical, or merely possible. I set these issues of formulation aside to focus on a more basic problem.

To say that something is constitutive of something else is, at the very least, to make a claim of metaphysical necessity: it is to say that one could not obtain without the other. But understood this way, proposals (i) and (ii) are implausible for a simple reason. Plausibly, animals and young children have mental states and perform intentional actions yet lack the capacity for introspective knowledge.[60] So it seems that introspective knowledge is not necessary, and so not constitutive, of these phenomena.

Two replies are possible. One is to bite the bullet and deny that animals and children genuinely have mental states and perform intentional actions. However, this is a highly revisionary move, at odds with common sense. It should be regarded as a last resort.

The other is to restrict the scope of (i) and (ii) to normal adult human beings. On this line, it would be constitutive of *normal adult human* minds and agency that mental states are introspectively knowable. This avoids the counterexamples, but at a cost: it is unclear what remains of the constitutivist claim. It is hard to make sense of something's being constitutive only for a subclass of a phenomenon: it is bizarre to say that it is constitutive of triangles *in the northern hemisphere* that their angles equal 180°. Saying that it is constitutive of normal adult humans that their mental states are introspectively knowable looks less like a constitutive truth and more like a psychological generalization.

---

[58] This formulation is found Shoemaker (2009: 35) and Smithies (2019: 157-158).
[59] This is reminiscent of various claims that Bilgrami makes (2006: 88-99).
[60] Burge ([2011] in 2013: 192) makes this point against constitutivist theories.

Since neither reply to the basic problem is palatable, I leave (i) and (ii) behind and continue the search for more promising formulations of constitutivism. Advocates of constitutivism can be presumed to defend more refined versions of the view.

The discussion of the restriction naturally leads into a different family of proposals, whose plausibility is less straightforward to assess.[61] Let *C-conditions* include possession of concepts of mental states as well as rationality and intelligence at a normal human level. Many constitutivist discussions are framed as attempts to give a constitutivist explanation of a certain kind of covariance. Schematically, the phenomenon to be explained can be captured as follows:

> CO For some mental state kinds *M*, it is necessarily the case that for all subjects *S*, if *S* meets conditions *C*, then *S* is in a mental state of kind *M* iff *S* believes that *S* is in a mental state of kind *M*.[62]

Three immediate comments. First, CO is underspecified: it does not say which kinds of mental state are covered. Second, it may be that CO would be better phrased in terms of 'being in a position to believe' rather than outright belief. Third, further refinements of the C-conditions may be needed to capture the exact boundaries of the covariance. With these caveats noted, CO points towards the sort of covariance constitutivists wish to explain. But what form might such an explanation take?

A first idea is to claim that belief about being in a mental state is constitutive of being in that state. Wright (1989) offered some influential programmatic remarks along these lines. His starting point is that in some domains, the necessary covariance between facts and true judgments does not arise because our judgments reflect independent facts, but because the judgments themselves, in some sense, determine how things stand. Color is often used as an illustration. On a familiar view associated with Locke, whether an object is red is not entirely independent of our judgments, but is partly fixed by what competent observers would say under ideal conditions. On this view, color judgments are (in Wright's terminology) "extension-determining" rather than "extension-reflecting" (1989: 246). (Wright does not himself endorse this view of color; his point is that it shows how judgments can, in principle, be determining rather than merely reflective.)

[61] See e.g. Bilgrami (2006: 1-45), Coliva (2016: 163-166 & further discussion in chap. 7 & 8), Wright in Bar-On & Wright (2024: 96-100).
[62] This is formulation is adapted from Coliva (2016: 164) and Wright in Bar-On & Wright (2024: 97)

By analogy, Wright suggests that some kinds of mental states may be like this. Whether I intend to *A,* for example, is not a fact prior to and independent of my judgments; rather, it is constituted by what I would judge under ideal conditions about my intention.[63]

Now, this sort of view clearly supplies a possible explanation of why CO holds. If, for some kinds of *M*s, being in *M* just is believing that one is in *M*, when the *C*-conditions obtain, then *S* being in *M* and *S* believing that *S* is in *M* cannot come apart. The necessary co-occurrence is guaranteed. And with it, we have a possible explanation of self-intimation.

This brings us to why the proposal is difficult to assess. If it is intended to explain special access, privileged access and directness, it plainly falls short as stated. These are all claims about knowledge, and the present constitutivist explanation of CO says nothing about knowledge. For all it shows, a subject's beliefs about being in *M* may always be true yet still fall short of knowledge. Perhaps this gap can be bridged, but the proposal is incomplete as it stands: it owes us an account of how extension-determining belief amounts to knowledge.

There is another way to develop the proposal. On this alternative development, the proposal is not meant to explain knowledge at all, but to displace the very project of providing an epistemology of introspection.[64] The absence of explanations of privileged or special access is not a defect, but the mark of a different conception of mind. This position calls into question the very framework of explananda that has guided the discussion so far. Since the implications are far-reaching, and since my aim in this introduction is to engage with the standard views that shape most of the current debate, I will set aside this revisionary proposal.

This brings us to another, less revisionary and at first sight more promising, way of explaining something in the vicinity of CO – one we find in Shoemaker. Shoemaker's framework is a kind of analytic functionalism: the project of providing functional definitions of kinds of mental states on the basis of *a priori* conceptual analysis (2003 [1981]: 272).[65] On this view, our concepts of mental states are to be analyzed in terms of their characteristic causes and

[63] Wright (1989), and the papers collected in (2001), develops these remarks in close connection with his interpretation of Wittgenstein's *Philosophical Investigations*. Wright himself does not endorse the view, and he does not determinately suggest that Wittgenstein would either. Rather, the view represents a position with Wittgensteinian pedigree, rather than one that Wright or Wittgenstein straightforwardly affirms.

[64] Coliva (2016: 163–165) discusses and offers some considerations in support of a version of this view. What I miss in that discussion is further reflection on the parallel with color: even if color is partly dependent on our cognitive constitution, it does not follow that I cannot know that an object is blue. Why should judgments about mental states be different?

[65] It contrasts with 'Psychofunctionalism' which turns to empirical science for functional definitions.

effects. To illustrate, consider a toy functional analysis of the concept *pain*: something is pain iff it tends to be caused by bodily damage and tends to cause anxiety, screaming, crying, and the desire to be out of the state. This analysis could be refined in numerous ways, but it conveys the general idea.

Shoemaker does not provide many full analyses, but he insists that analyses like this must be amended in a crucial respect. For creatures with a certain level of sophistication, pain also characteristically has a:

> tendency to produce in the person the belief that there is something organically wrong with him (e.g. that he has been cut or burnt), [and also a] tendency to produce qualitative beliefs in the person, i.e. to make him think that he has a pain having a certain qualitative character (one that he dislikes) (2003 [1975]: 189).[66]

Three points are worth stressing. First, Shoemaker's claim is that it is part of our concept *pain* that pain has a tendency to produce introspective belief in creatures that meet conditions *C*.

Second, this claim about *a tendency* to give rise to introspective belief in creatures with *C* is weaker than CO (a thesis about actual belief) and so cannot explain CO as stated. However, as stressed above, the formulation of CO requires refinement, and a natural way of doing so is to rephrase it in terms of being in a 'position to believe'.[67] And here 'tendencies' is apt.

Specifically, in speaking of tendencies, I take Shoemaker to be making a claim about a *capacity* for introspective belief – as contrasted with *actual* introspective beliefs. It is a general truth that my having a capacity to *A* does not entail that I actually *A*; capacities can go unexercised. By parity, one can have a capacity for introspective belief, without exercising it.[68]

Third, Shoemaker's claim about pain can be generalized. For instance, it is part of our concept *itch* that itches have a tendency to produce introspective belief in creatures with *C*. Similarly, it is part of our concept *belief* that beliefs have a tendency to produce introspective belief in creatures with *C*, and so on for other kinds of introspectively knowable mental states.[69]

From these points we can derive the thesis CAPACITY:

---

[66] Shoemaker says that each person has "'direct' access to this qualitative character" through "introspection. And introspection, whatever else it is, is the link between a man's mental states and his beliefs (or his knowledge or awareness of) those states" (2003 [1975]: 189).
[67] This reformulation is also encouraged by the discussion of ignorance and fallibility in §3.
[68] For instance, I have the capacity to punch people in the face, but I never exercise it. I am not violent.
[69] See Shoemaker (1999: 740; 2009: 36).

> CAPACITY It is part of our concepts for certain kinds of mental states that, necessarily, when a creature *S* meets conditions C*, S* has a capacity for introspective belief that she is in those states.

Shoemaker never explicitly embraces CAPACITY, but it is natural to ascribe that thesis to him. On this view, to have *C* – to be rational, concept-possessing, and intelligent at a normal human level – just is to be the kind of creature for whom pains (and other mental states) confer the capacity for introspective belief. In this sense, our having the capacity is guaranteed by, and explained by, our possession of *C*. That makes CAPACITY a constitutive thesis in a straightforward sense: necessarily, meeting conditions *C* grounds the very possibility of introspective belief.[70]

Some immediate motivation for CAPACITY comes from the considerations of self-blindness (see §4.1 again) – the condition in which a person meets conditions *C* but lacks a capacity for introspective belief. If CAPACITY is true, it immediately explains why self-blindness is impossible. If meeting the *C*-conditions necessarily bestows a capacity for introspective belief, then a creature with *C* but without a capacity for introspective belief is impossible. In this way CAPACITY vindicates Shoemaker's longstanding claim that self-blindness is not a genuine possibility.

Unlike the more revisionary extension-determining version of constitutivism considered above, CAPACITY does not conflict with the epistemological project of explaining how introspective beliefs amount to knowledge. However, as stated, CAPACITY leaves the central explananda untouched: it says nothing about knowledge. Developing the view into a full theory is an ambitious task that lies beyond the scope of this discussion. But to indicate how such a development might proceed, and where it is likely to encounter difficulties, I close with some general remarks.

With respect to directness constitutivism faces no major obstacle. On CAPACITY, introspective belief does not arise from inference, evidence, or reasons, so the condition of directness is naturally satisfied. The same goes for self-intimation: all that is required is an additional story about how mental states become available to introspection. A natural suggestion is that availability is explained by some triggering mechanism – for instance, a state becomes available when the agent reflects on whether she is in it.[71]

---

[70] Shoemaker offers various further characterizations of this constitutive connection. Sometimes he describes it in terms of *identity* (1996: 244), other times he appears to describe something like *entailment* (1996: 243), and occasionally he makes claims in terms of terms of *essences* (1996: 27) and *part-whole* relations (2009: 42).

[71] Shoemaker offers scattered remarks on the required mechanism at various places (1996: 91–93, 241–242; 2009: 48–50) but is adamant that he is not offering an account of it.

By contrast, matters are much less straightforward for privileged access and special access. To explain privileged access, two substantial additions are needed. First, the view must show why introspective beliefs, as construed by CAPACITY, are more secure in delivering true beliefs about mental states than perception or testimony. Nothing in CAPACITY itself explains this. Second, it must be supplemented with an account of introspective warrant and an explanation of why introspection yields knowledge at all. Constitutivism is compatible with a range of epistemological theories, internalist and externalist, and its explanatory resources will depend heavily on how it is paired.

For special access, the problem is to explain the necessary exclusivity of introspective knowledge. Suppose that my twin and I share a pain. What, on the constitutive story, ensures that the constitutive tie links the pain only to my introspective belief that I am in pain, and not also to beliefs about my twin's pain? What explains that I cannot introspectively know that my twin is in pain? Without answers, special access remains unaccounted for. These remarks do not show that no answers are available. But they do reveal that further work is required before CAPACITY can be said to meet the explananda.

To take stock: I have distinguished several constitutivist proposals. The most straightforward versions proved implausible, while more programmatic suggestions are difficult to assess. Shoemaker's version, reconstructed here as CAPACITY, holds promise because it explains why self-blindness is impossible. But CAPACITY on its own does not explain knowledge, and whether it can ultimately deliver explanations of special access, privileged access, and the rest depends crucially on the epistemological theory of warrant with which it is paired. In its present forms, constitutivism remains incomplete, but it is still a serious option.

§4.6. Summary

I have now surveyed the main theories of introspective knowledge. Each makes some progress on the explananda but falls short in different ways.

In §§4.1–4.2, I examined inner-sense and acquaintance theories, which treat introspection on the model of observation. They fail to explain privileged and special access, give no adequate account of self-intimation, and raise further puzzles: if introspection is inner observation, why does it lack its own phenomenology, and why are introspective capacities not subject to loss in the way other observational capacities are?

In §4.3, I turned to consciousness-based theories such as Peacocke's. These may capture privileged access and directness, but not special access, and in several cases they leave self-intimation unexplained by not offering a plausible story for introspective knowledge of non-phenomenal states.

In §4.4, I examined transparency theories. Some versions show promise across all four explananda – privileged access, special access, directness, and self-intimation – but a scope problem remains: it is unclear whether transparency can be extended beyond belief to cover the full range of mental states.

In §4.5, I assessed constitutivist theories. I suggested that Shoemaker's version is attractive since it explains why self-blindness is impossible. However, it remains incomplete: its success depends on further details.

To conclude: no existing theory is fully satisfactory. Yet the survey also shows where progress seems possible. Transparency theories and Shoemaker-style constitutivism suggest that a satisfactory account of introspective knowledge may be within reach, provided that their core insights can be developed in a way that meets the explananda.

Throughout this survey I have assumed monism (see §1): that there is a single capacity for introspective knowledge, and that the theories under review are to be evaluated by how well they model that capacity. This evaluative framework is standard in the literature, but the sort of shortcomings I have highlighted have led some to take pluralism as a serious alternative. This is further discussed in Paper II.

I have surveyed several theories of introspective knowledge, evaluating their prospects for explaining familiar epistemic features such as privileged access, special access, directness, and self-intimation. I now turn to another important phenomenon that has often been discussed in connection with introspection: *immunity to error through misidentification* (IEM).

§5. Immunity to Error through Misidentification

At a first approximation, a judgment exhibits error through misidentification (EM) when its agent has issued the judgment on the basis of a misidentification. A judgment is *immune* to such error (IEM) when this kind of error is impossible, and *vulnerable* (VEM) when such error could occur. Following Shoemaker (1968), most theorists hold that all introspective first-person judgments are IEM.[72]

Two immediate clarifications will help frame my discussion. First, following Pryor (1999), it is customary to distinguish two kinds of misidentification errors, both concerning judgments of the form 'a is F'.[73] On the first, which

[72] A notable exception is Cappelen & Dever (2013: 130-133).

[73] See his characterization of *de re*-misidentification (1999: 274) and *wh*-misidentification (1999: 282). My formulation of these two kinds of error makes certain simplifications. Pryor characterizes the first in terms of propositional justification for believing an identity claim of the form 'a = b' (1999: 274), and the second in terms of his theoretical notion of "grounds that offer knowledge" (1999: 281). I have set aside these here. The first, and most important reason

will be my focus, *S* errs through misidentification when *S* judges *that a is F* as a result of having misidentified some object, b, as a. On the second, *S* errs by moving from knowledge *that something is F* to the mistaken judgment *that a is F* because *S* wrongly takes a to be the relevant witness of the existentially quantified claim. While this second phenomenon is interesting, I set it aside here, since it seems to me that the first kind captures something especially important about introspection. In what follows, my use of the terms 'IEM', 'EM', and 'VEM' is reserved for misidentification errors of the first kind.

Second, in saying that IEM captures something important about introspection, I do not mean to suggest that IEM is distinctive of introspective judgment. It is not. There is a substantial debate about the scope of the IEM phenomenon – for example, whether memory-based judgments are also IEM. But I set that debate aside and focus on introspection.

To appreciate what is at stake, let us look closer at the phenomenon of identification itself. The clearest illustration comes from ordinary perception. Consider a stereotypical case of perceptual knowledge: I acquire knowledge that John is bearded on the basis of seeing him. This piece of visual knowledge depends crucially on my ability to recognize John as John. The exact workings of this ability may be complex, but the important point here is simple: in recognizing John, I form my belief on the basis of his visible characteristics. John looks a certain way, and his looks allow me to visually identify him.

Looks can deceive. If conditions conspire against me, or if I encounter someone who closely resembles John, I may mistake that person for John. In case I do mistake that person for John, my judgment that John is bearded is EM, precisely because I have misidentified someone as John and then judged that John is bearded on that basis. More generally, perceptual judgments of the form 'John is F' are VEM.

It is against the background of this understanding of identification that we should understand Shoemaker's claim (1968; 1996; 2011) that introspective first-person judgments – e.g. the judgment that *I* am in pain – are IEM.[74] The kind of error that can afflict my perceptual judgment that John is bearded simply cannot occur in the case of introspective judgment that I am in pain. This naturally invites the question: what explains that first-person introspective judgments are IEM? Three different accounts merit consideration.

---

for present purposes, is presentational: introducing such machinery would overburden the discussion at the level of generality I am aiming for. However, there is also a second, substantive, point: the phenomenon that theorists in the tradition of Shoemaker (1968) seek to capture in terms of 'error through misidentification' does not, or so it seems to me, require that one possess justification. What matters is whether one in fact identifies someone as someone – not whether that identification is backed by justificatory grounds.

[74] This formulation raises questions about whether introspective judgment that Carl is in pain is IEM too. I set that aside.

The first is the ASYMMETRIC ACCESS explanation:

> ASYMMETRIC ACCESS Necessarily, for all subjects *S*, whenever *S* issues an introspective first-person judgment, that judgment is IEM because, for every subject *x* distinct from *S*, it is impossible for *S* to introspect *x*'s mental states.[75]

The underlying idea is simple: since I cannot introspect the mental states of others, I cannot introspectively mistake someone else's mental states for my own. This implies that my introspective judgment that I am in pain cannot be EM.

At first sight, ASYMMETRIC ACCESS has considerable plausibility. However, on closer inspection it faces difficulties. Consider shared mental states. Suppose that you and I share a pain, I feel the pain and judge that I am in pain. If so, I have introspected a mental state that is not uniquely mine. This contradicts ASYMMETRIC ACCESS, which claims that it is impossible to introspect another's mental states. So, ASYMMETRIC ACCESS fails to explain why introspective judgments are IEM.[76]

Three clarifications are in order.

First, one might try to salvage ASYMMETRIC ACCESS by adding a qualification: *so long as subjects do not share mental states*, it is impossible for one to introspect another's states. This would block the counterexample. But such a repair is *ad hoc*. The spirit of ASYMMETRIC ACCESS is the claim that introspection is necessarily confined to one's own mentality. Stipulating away shared mental states deprives the principle of explanatory force.

Second, one might object that whether shared mental states refute ASYMMETRIC ACCESS depends on which theory of introspection is correct – perceptual, transparency-based, constitutivist, or otherwise. In response, I grant that a final verdict will depend on the theory adopted. My point is that if we consider introspection as ordinarily understood – our familiar way of forming first-person beliefs about our own mental states – nothing about it rules out introspective belief in cases of shared mental states. If something blocks this possibility, something other than introspection (in our ordinary understanding) must do so. Until we have a story of what that would be, ASYMMETRIC ACCESS is under pressure.

Third, saying that it is possible to introspect another person's states in cases of shared mental states does not amount to denying that we enjoy special

[75] For endorsement of ASYMMETRIC ACCESS, see Pryor (1999: 279-280, 282).

[76] Notice that shared mental states do nothing to put doubt on the claim that introspective first-person judgments are IEM. If I am in pain, I share this pain with someone else, and I introspectively judge that I am in pain, my judgment is not mistaken.

access to our own mental states. Special access concerns a distinctive way of knowing, not the mere reach of introspective belief. My claim here is only that ASYMMETRIC ACCESS mischaracterizes the limits of the latter.

With these objections addressed and the main clarifications in place, the case against ASYMMETRIC ACCESS stands. We have good reasons to reject it.

A second explanation is the INFALLIBLE IDENTIFICATION explanation:

> INFALLIBLE IDENTIFICATION Necessarily, for all subjects *S*, whenever *S* issues an introspective first-person judgment, that judgment is IEM because *S*'s introspective self-identification cannot be mistaken.

Again, the thought is straightforward: if introspection necessarily involves correct self-identification, then introspective EM is impossible since EM requires a misidentification.

However, there are two serious problems.

First, INFALLIBLE IDENTIFICATION lacks motivation and seems implausible. All our other identificatory capacities are fallible; why should introspection alone be the exception? Without independent reason to believe in such infallibility, INFALLIBLE IDENTIFICATION lacks plausibility.[77]

Second, and more fundamentally, the explanation is at odds with the way we issue introspective first-person judgments. Shoemaker (1968), drawing on Hume, P.F. Strawson and Wittgenstein, emphasizes that when one self-ascribes mental properties, one does not rely on self-recognition.[78] Strawson puts the point well:

> When a man (a subject of experience) ascribes a current or directly remembered state of consciousness to himself, no use whatever of any criteria of personal identity is required to justify his use of the pronoun "I" to refer to the subject of that experience. It would make no sense to think or say: *This* inner experience is occurring, but is it occurring to *me*? (This feeling is anger; but is it I who am feeling it?) […] There is nothing that one can thus encounter or recall in the field of inner experience such that there can be any question of one's applying criteria of subject-identity to determine whether the encountered or recalled experience belongs to oneself – or someone else (1966: 165).

Two claims are made here, each reinforced by a contrast with perception.

First, ordinary mental self-ascription makes no use of criteria of self-identification: I do not first determine whether I am the person in pain and only then judge that I am in pain – I simply judge that I am in pain.

---

[77] For further discussion, see Wright (2012: 253).

[78] Shoemaker refers to the passage from Hume quoted in §2 above, Strawson (1959: 107-111) and Wittgenstein (1958b: 66).

Second, such self-ascriptions are epistemically appropriate ('justified' in Strawson's terminology). When I self-ascribe pain in this way, I can, and typically do, introspectively *know* that *I* am in pain. This knowledge is "identification-free".[79]

In both respects this contrasts with perceptual judgments: perceptual judgment that John is bearded typically relies on identification, and whether it qualifies as knowledge depends in part on such identification.

Hume's observation (again see §2) reinforces these two claims. Phenomenology does not disclose the existence of a self: there is no appearance, no 'look' or 'feel', on the basis of which a subject could identify itself as the bearer of its own mental states.[80] For this reason, there is nothing on which an identification – fallible or infallible – could be based. Without such grounds, the idea of an infallible identification is a non-starter.

Taken together, the observations suggest that INFALLIBLE IDENTIFICATION is fundamentally mistaken. Introspective judgment does not involve identification at all, let alone of an infallible kind.

Strawson's and Hume's remarks naturally motivate Shoemaker's own explanation:[81]

> NO IDENTIFICATION Necessarily, for all subjects *S*, whenever *S* issues an introspective first-person judgment, that judgment is IEM because *S*'s judgment involves no identification.

Again, the idea is simple: EM requires misidentification, and where there is no identification, misidentification is impossible. Introspective first-person judgments involve no identification. And that is why all introspective first-person judgments are IEM.

What objections can be pressed against NO IDENTIFICATION? Unlike the two other explanations considered, there seems to be no obvious counterexamples to NO IDENTIFICATION.[82] Instead, it faces a recurring complaint:

---

[79] This term is from Evans (1982: 242f).

[80] Shoemaker phrases it in the following terms: introspection "does not involve being presented to oneself as an object" (1968: 563).

[81] Numerous theorists endorse some version of it: see Bar-On (2004), Coliva (2006), Evans (1982), Morgan (2012), Shoemaker (1968; 2011), Wright (2012) and both Bar-On and Wright (2024: chs. 7.2. and 8.2).

[82] However, Pryor offers a problem case from which he concludes that "The fact that one's belief is identification-free does not show that the belief is immune to errors of misidentification" (1999: 292). On his use of terms, this renders NO IDENTIFICATION false by implication. But his use of 'identification' is stipulative, equating it with having justification to believe a proposition of the form 'a = b'. As noted earlier (n. 73 above), I take this to mischaracterize the relevant phenomenon at issue. One can misidentify without having the justification Pryor

without a fuller account of identification, it is unclear what NO IDENTIFICATION asserts.[83] What is identification such that its absence guarantees IEM and its presence guarantees VEM? And what is introspection like, such that all first-person judgments it yields are identification-free in the relevant sense? Until these questions are answered, NO IDENTIFICATION can seem underdeveloped.

This complaint has some force. As the literature on IEM has grown increasingly technical, it has also become increasingly unclear what the core phenomena are.[84] What is identification, really? Still, Shoemaker's work – though gnomic in other respects – isolates a relatively clear and recognizable phenomenon (as articulated in the remarks by Strawson and Hume). Although this phenomenon sometimes gets buried in the literature, it gives NO IDENTIFICATION enough intuitive content to be both intelligible and compelling.

To repeat the relevant contrast in simple terms: perceptual knowledge that John is bearded involves perceptual identification of John, a process that makes use of surface characteristics. Introspective knowledge that I am in pain involves no analogous process. That is surely not the whole story about identification, but it is enough to answer the basic complaint against NO IDENTIFICATION. What is absent in introspection – and what therefore explains its immunity – is precisely the kind of identification familiar from perception.

However, two further challenges stemming from the recurring complaint remain.

The first challenge concerns extensions. A general explanation of IEM should apply across all cases. Consider demonstrative perceptual judgments. As Shoemaker notes, when I simply point at an object and assert 'this is red,' I may be wrong about the object's redness but I cannot have misidentified some other object as the one I demonstratively pick out with my use of 'this' (1968: 558).[85] Such demonstrative perception-based judgments are IEM.

The immediate difficulty for NO IDENTIFICATION is obvious but uninteresting: being confined to introspection, it cannot explain perception-involving cases of IEM. However, a more interesting difficulty arises once we consider the more general explanatory claim from which NO IDENTIFICATION

speaks of, and one can have such justification without identifying. Since Pryor speaks of something different, it would seem that his remarks do not falsify NO IDENTIFICATION.

[83] See Garcia-Carpintero (2018) and Cappelen & Dever (2013: 136-137) where this sort of complaint comes to the fore.

[84] For a recent overview of the field, see Coliva & Palmira (2025): 'identification' is used for a range of different things, not all of which seem related or equally relevant.

[85] More carefully, the judgment expressed by 'this is red' is IEM when it is (i) made on the basis of perception, (ii) 'this' picks out an object demonstratively, and (iii) the use of 'this' is not determined by an intention to speak or think of an object one has encountered before.

derives: that necessarily, whenever a judgment of the form 'a is F' is not based on identification, it is IEM.

Here a worrying disanalogy appears to arise. In the perceptual-demonstrative case, a fully visible object is directly referred to with 'this' without prior identification, and the judgment (expressed by) 'this is red' is IEM.[86] In the introspective case, 'I' is applied without any phenomenological presentation of the self, and the judgment (expressed by) 'I am in pain' is IEM. But these two seem to be very different ways in which identification can be absent. What unifies them? Unless we can say more, it is unclear whether NO IDENTIFICATION, or the general 'absence of identification' idea behind it, can provide a uniform explanation of IEM.

The second challenge concerns first-person introspective knowledge. Even if NO IDENTIFICATION is the correct explanation of why introspective first-person judgments are IEM, it threatens to make the explanation of *introspective first-person knowledge* mysterious. Shoemaker presses this point:

> it can seem puzzling that self-awareness [read: introspective knowledge], of the sort we are concerned with, does not involve being presented to oneself as an object. I cannot see the redness of a thing without seeing the thing that is red, and it would seem that it should be equally impossible to be aware of a state of oneself without being aware of that which has that state, i.e. oneself (1968: 563).

The puzzle is this. If phenomenology presents me only with a succession of mental properties – pain, tickle, itch – without disclosing the existence of a bearer, it would seem to warrant at most existential judgments: that there is pain, that there is a tickle, and so on. Yet introspection delivers more: I introspectively *know* that *I* am in pain. The challenge, then, is to explain how such first-personal knowledge is possible when it involves no self-identification. And unless NO IDENTIFICATION can be paired with a plausible account of this knowledge, it risks standing in unresolved tension with the existence of first-person introspective knowledge.

This is a challenge and not a refutation. Indeed, although I do not explicitly frame it in those terms, Paper I develops a response. There I argue that first-person knowledge is indispensable for certain rationally required projects. This indispensability grounds its status as basic knowledge, and so explains how I can have introspective knowledge that *I* am in pain without identification. If that explanation succeeds, NO IDENTIFICATION can be retained.

[86] Articulating this observation in detail is hard. One way to put it: I do not think of this object first, and then apply a concept corresponding to 'this', as if the object were the unique *this* in the circumstances.

The present discussion clarifies why IEM merits separate treatment in this introduction. The central thread is this: introspective first-person judgments are IEM, and this fact calls for explanation. Yet, the explanation appears unexceptional, since it has nothing specifically to do with introspection: IEM arises from the general fact that some judgments are made without identification. But what remains striking is that introspective judgments – made in an identification-free way – can *yield* knowledge. How? In perception, identification seems necessary for knowledge that John is bearded, whereas in introspection it clearly is not required for knowledge that I am in pain.

Here lies the important philosophical task: to explain how identification-free knowledge is possible. IEM marks the point where the familiar features of introspection – privileged access, special access, directness, self-intimation – intersect with a deeper puzzle about how identification-free judgments can amount to knowledge.

However, as noted in §1, there is another reason to linger on identification and misidentification. Ascriptions of (mis)identification exhibit the peculiar asymmetry phenomenon. Pryor illustrates it in his seminal discussion of IEM. He invites us to imagine that "I see a person in the mirror with a black eye [and I] misidentify him as myself", constituting a case where "I misidentify someone else as myself" (1999: 276). However, Pryor notes that "I can also misidentify someone who is in fact me as someone else", presenting the following case in support: "I see myself in the mirror wearing a yellow hat, and mistake myself for George" (1999: 276).

What is peculiar? It is natural to think of misidentification as a false *identity judgment*, i.e. a judgment embedding something like the proposition that George = Jim. But while identity is symmetric, the sentences 'I misidentified myself as George' and 'I misidentified George as me' are not equivalent: their truth-values can diverge. This raises deep questions in natural language semantics. How should we explain the truth-value asymmetry we see in this case? That is the topic of the next section (§6) which serves an introduction to the topic explored in Paper III: asymmetric attitude reports.

§6. Asymmetric Attitude Reports

The asymmetry phenomenon has been recognized in the philosophy of language, though typically only piecemeal. Few discussions have considered the full range of relevant data, and fewer still have attempted a systematic explanation. In what follows, I present the data and provide some background for theorizing about them.

Stalnaker discusses the following two sentence pairs (1999: 128-129):

(1a) Daniels believes that the man in the faded denim leisure suit is David Kaplan.
(1b) Daniels believes that David Kaplan is the man in the faded denim leisure suit.
(2a) Daniels believes that I am David Kaplan.
(2b) Daniels believes that David Kaplan is I.

Stalnaker observes that despite the symmetry of identity the b-sentences differ in meaning from the a-sentences (1999: 129).

Saul Kripke makes similar remarks about the following pair:

(3a) I mistook the President of Harvard for a teaching fellow.
(3b) I mistook a teaching fellow for the President of Harvard.

He notes that (3a) is naturally understood as true when the President is taken for a teaching fellow – perhaps due to his youthful appearance (2005: 1029). By contrast, (3b) is naturally understood as true when "some teaching fellow one encountered was so old and distinguished looking" that he was mistaken for the President of Harvard (2005: 1029).

The asymmetry phenomenon extends beyond reports of belief and mistake. Delia Graff Fara (Ms.) presents the following pair of desire-reporting sentences:

(4a) I want to be Marlene Dietrich
(4b) I want Marlene Dietrich to be me.

Intuitively, (4a) seems to be true when I want something to happen to *me* – that I become Marlene Dietrich – and (4b) seems to be true when I want something to happen to *Marlene Dietrich* – that she becomes me. Analogous contrasts appear with imagination reports:

(5a) I imagined being Napoleon.
(5b) I imagined Napoleon being me.[87]

As these cases suggest, the asymmetry phenomenon arises whenever an attitude verb embeds what appears to be an identity statement of some sort. To situate these observations, I turn to the tradition of theorizing about attitude reports, and research on the semantics of the copula.

[87] Observe that (4a) and (5a) both involve PRO: an unpronounced pronoun at the level of logical form with a mandatory *de se* reading (see Chierchia 1989): I, the attitude-holder, think of myself as myself.

Frege's work is a natural starting point. In *Über Sinn und Bedeutung* (1892), Frege famously distinguished between the *Bedeutung* ('reference') and *Sinn* ('sense') of terms to explain how sentences of the form 'a = a' and 'a = b' can have different cognitive significance, even when 'a' and 'b' co-denote. A person who knows that the Evening Star = the Evening Star can still learn something new when informed that the Evening Star = the Morning Star, because although the reference is the same, the cognitive significance differs.

This framework was extended to belief-sentences, showing how the following can both be true:

(6a) Hammurabi believes that Hesperus = Hesperus.
(6b) Hammurabi does not believe that Hesperus = Phosphorus.

Although 'Hesperus' and 'Phosphorus' both refer to Venus, Hammurabi associates different senses with the names. This offers an explanation how he can believe (6a) without also believing that Hesperus = Phosphorus. This gave rise to the now familiar lesson that co-referential terms cannot always be substituted *salva veritate* in attitude contexts.

With Frege began an enormously influential research program in the philosophy of language. A vast literature grew on the semantics of belief and the semantics of names and other singular terms.[88]

Not long after Frege, Russell made an equally influential remark on the copula 'be' in *Introduction to Mathematical Philosophy*:

> The proposition "Socrates is a man" is no doubt *equivalent* to "Socrates is human", but it is not the very same proposition. The *is* of "Socrates is human" expresses the relation of subject and predicate; the *is* of "Socrates is a man" expresses identity. It is a disgrace to the human race that it has chosen to employ the same word "is" for these two entirely different ideas – a disgrace which a symbolic logical language of course remedies (1919: 172).

This passage is remarkable in many ways.[89] But what is most important for present purposes is the explicit recognition of two distinct uses of the copula:

[88] Moreover, identity itself has remained a central philosophical topic.

[89] Russell's use of quotation marks for propositions and italics for words is somewhat unusual today. Substantively, his claim (in different terms and notation) that the sentences 'Socrates is a man' and 'Socrates is human' express distinct propositions although they are "no doubt equivalent" is a little puzzling. If the equivalence is merely material, the remark is trivial; if it is necessary, it is not immediately clear how the propositions can fail to be "the very same". In addition, in contrast to what Russell says, the sentence 'Socrates is a man' is arguably not an identity sentence at all. For instance, it is not reversible ('a man is Socrates' makes little sense). This shows only that the example is badly chosen, not that the distinction Russell was aiming to capture is mistaken. Finally, his remark about a "disgrace to the human race" is rhetorically memorable but philosophically inessential. However, the suggestion that the ambiguity of 'be'

'be' as predicative and 'be' as identity. Once this distinction had been drawn, the semantics of 'be' received comparatively little sustained philosophical attention – at least until relatively recently, when it has re-emerged as a central topic.[90] During this period when the copula received relatively little sustained attention in the philosophy of language, it became the focus of intensive research in linguistics, especially from the 1970s onward. That work introduced fine-grained distinctions between different copular constructions, some of which are directly relevant for understanding the asymmetry phenomenon.

Roger Higgins (1979) made an important contribution by distinguishing four types of copular sentences:[91]

SPECIFICATIONAL	(7) The President of the U.S. is Joe Biden.
IDENTIFICATIONAL	(8) That is Dan.
EQUATIVE	(9) Cicero is Tully.
PREDICATIVE	(10) Cicero is hungry.

Equative and predicative sentences are familiar to philosophers. By contrast, specificational and identificational sentences may be less so. Roughly, specificational sentences identify which object satisfies a given description. Identificational sentences, by contrast, serve to introduce or pick out an individual for an audience.

Researchers in linguistics have investigated whether the various uses of 'be' across predicative, equative, specificational, and identificational sentences reflect a single underlying element or several distinct ones. A central issue in this literature is whether the copula has one uniform syntax and semantics, or whether the different sentence types distinguished above involve 'be' with different syntactic and semantic properties.

Three broad schools of thought have emerged.

One school of thought claims that 'be' at bottom always has the same syntax and semantics (den Dikken 2006; Moro 1997; Partee 1986; 2010: 30; Williams 1983), one captured by PREDICATION:[92]

---

is a defect of ordinary language to be remedied by symbolic logic probably reflects Russell's primary interest in his discussion: it is not natural language as such, but the construction of a logically perspicuous ideal language.

[90] Renewed philosophical interest in the semantics of 'be' arose after Hofweber (2005) and his discussion of sentences such as 'The number of planets is 9' (for earlier discussion, see Quine 1960: 195-200).

[91] For a more recent overview, see Mikkelsen (2011).

[92] One potential difficulty for this view deserves mention. It assimilates equatives such as 'I am Carl' to predication. Partee (1986) introduces a type-shift rule (*ident*) whereby the phrase 'is

PREDICATION [[be]] = λPλx. Px.

In this function, 'be' takes a predicate 'P' and a singular term 'x' and returns the true iff the referent of 'x' has the property expressed by 'P'.

Other theorists maintain that at least two denotations are required to handle equative, predicative, specificational and identificational sentences, adding IDENTITY alongside PREDICATION:

IDENTITY [[be]] = λyλx. x = y.

IDENTITY takes singular terms 'y' and 'x' and outputs the true iff the referent of 'x' is numerically identical to the referent of 'y'. IDENTITY is standardly taken to provide the denotation for (9) and equative clauses more generally.

Still others argue for more than two denotations. Romero (2005), building on Sharvit (1999), posits a distinct denotation for 'be' in specificational sentences,[93] whereas Moltmann (2012; 2013) introduces a separate denotation for 'be' in identificational sentences.[94]

When we survey the field, something of a disciplinary contrast emerges. Philosophers of language, following Russell, tend to treat 'be' as ambiguous between IDENTITY and PREDICATION, and tend to treat this distinction as sufficient for semantic analysis. By contrast, linguists have reached no such consensus. At present, it remains an open question how many denotations 'be' requires – one, two, or several. The natural language data are complex and varied, and different classes of copular sentences appear to resist uniform treatment.

While there are many data points of interest, I will limit discussion to those most important for our purposes. Like the asymmetry phenomenon, these are data that pre-theoretically pattern with identity sentences – they have an identity-ish feel – but where 'be' does not seem to behave like numerical identity. These cases suggest that the link between 'be' and IDENTITY is less straightforward than philosophical orthodoxy has assumed, and that 'be' can have a different denotation even when flanked by referential terms.

Carl' is type-shifted into a predicate (λx. x = Carl) (i.e. the property of being Carl). But on reflection, this is counterintuitive: whereas the predicative 'be' is meaningless, 'be' in equatives is not: intuitively, it expresses identity. The one-*be* view either leaves identity unaccounted for or collapses into a two-*be* theory in disguise.

[93] Romero's 'copula of specification' is SPEC $[[be_{spec}]] = \lambda x \lambda y \lambda w. [y(w) = x]$ (Romero 2005: 715), which equates a term of type <s,e>, an individual concept (a function from worlds to individuals), with a term of type <e>. For a competing suggestion, see Schlenker (2003).

[94] Moltmann claims that "Identificational sentences have a range of properties that make clear that they are neither identity statements nor subject-predicate statements of any sort" (2013: 347). Moltmann's 'identificational be' is reminiscent of Romero's SPEC. For discussion of competing suggestions on how to handle identificational sentences, see Mikkelsen (2011).

Consider identificational sentences such as (8) repeated below:

(8) That is Dan.

The reversed version of (8), namely (8 rev), is unacceptable in most contexts:

(8 rev) # Dan is that.

This sensitivity to NP order is not predicted if 'be' simply denotes IDENTITY. A similar point emerges if we attempt substitution with 'be identical to':

(8=) That is identical to Dan.

The resulting apparent shift in meaning suggests that 'be' is doing something different here than expressing identity. While (8=) is grammatical, the contexts in which we would use it seem different from the contexts in which we ordinarily use (8).

Counterfactuals exhibit a similar pattern:

(11) If I were Napoleon, Napoleon would be me.
(11=) If I were identical to Napoleon, Napoleon would be identical to me.

While it may not be entirely clear what (11) and (11=) say – they are rather puzzling – it seems that (11) can be false in certain contexts, whereas (11=) must be true, if it is true at all.[95]

Even so-called *theoretical identifications* (see e.g. Kripke 1981) show a similar pattern. Consider 'Water is H_2O'. If 'be' here is the 'be' of identity it is not immediately clear why the reversed sentence 'H_2O is Water' is less natural.

Taken together, these data show that 'be' does not always behave like IDENTITY, even in sentences that pre-theoretically look identity-like. This makes it natural to question whether appeals to IDENTITY can account for all the phenomena.

Having introduced some of the relevant data and theoretical perspectives on copular clauses, we now return more narrowly to the asymmetry phenomenon. On the view I argue for, the asymmetry phenomenon finds its explanation in the fact that the order of terms around 'be' in equative clauses makes a semantic contribution. Let the left-most noun phrase, NP, in a 'NP be NP' clause be *the precopular* NP, and the rightmost NP be *the postcopular* NP. On my account, equative clauses have a fixed structure: the precopular NP

[95] If it is false, it is presumably necessarily false.

introduces a bearer while the postcopular NP provides an identity attribution, supplying an implicit answer to 'Who is this?'.

To capture this, I introduce a novel denotation for 'be':

WH-IDENTITY [[be]] = λyλx. who x is is y.

WH-IDENTITY takes an ordered pair of expressions 'y' and 'x' of type <e> and yields the true iff who x is is y.

This order-sensitivity is directly reflected in simple question-answer (QA) pairs:

(12Q) Who is that? (Context: we see a woman approaching).
(12A) That is Joan Smith.

(13Q) Who is she? (Context: you have just talked of a woman using 'she').
(13A) She is Joan Smith.

These exchanges show how information is conveyed through the fixed structure of the answer: the pre-copular NP introduces the individual, and the postcopular NP specifies *who that individual is*. WH-IDENTITY makes sense of this pattern by assigning 'be' a denotation that encodes a certain directionality of the construction.

With this in hand, consider the somewhat more complex case of asymmetry data:

> *The Masked Ball*. At a masked ball, Katherine wears Rosaline's costume, Rosaline wears Maria's costume and Maria wears Katherine's costume. Their friend Biron knows to whom the costumes actually belong, but not that the switches have occurred. Biron sees the women enter. He mistakes Katherine for Rosaline, Rosaline for Maria, and Maria for Katherine.

(14a) Biron thinks Katherine is Rosaline.
(14b) Biron thinks Rosaline is Katherine.

At the *Masked Ball*, (14a) has a salient true reading, while (14b) has a salient false reading.

My explanation is this: (14a) attributes to Biron a belief about who Katherine is; (14b) attributes to Biron a belief about who Rosaline is. The two sentences differ precisely because WH-IDENTITY makes the order semantically significant. With 'be' understood as WH-IDENTITY, (14a) can be seen as an

answer to the question ‘Who does Biron think that Katherine is?’ and (14b) as an answer to ‘Who does Biron think that Rosaline is?’.

The same explanation extends to other asymmetric pairs, such as those involving ‘desire’ and ‘imagine’ (see (4a)/(4b) and (5a)/(5b) above): the asymmetry arises because the denotation of ‘be’ in the embedded clause is given by WH-IDENTITY.

In one sense, much of this should be familiar. We already know that equatives answer questions of the form ‘Who is x?’, and our ordinary practice reflects this: we say ‘That is Joan Smith’, not ‘Joan Smith is that’ or ‘Joan Smith is her’ in answer to ‘Who is she?’.

What is less familiar is my claim that this requires recognizing a new denotation for ‘be’. Indeed, one might suspect that WH-IDENTITY simply is a roundabout paraphrase of IDENTITY. When I say ‘I am Carl’, this serves as an answer to ‘Who are you?’. But it is of course also true that I am numerically identical to Carl. How, then, could the relations that WH-IDENTITY and IDENTITY encode come apart?

I do not offer a metaphysical account of their difference. However, my proposal is that the linguistic data strongly motivate the search for one. The data already surveyed ((8)-(13A)) suggest that ‘be’ and ‘be identical to’ do not express the same property: the two make different semantic contributions to the sentences in which they occur. That is why WH-IDENTITY deserves recognition as a distinct semantic contribution of the copula.

Consistent with the contrasts exhibited in (8)-(13A), similar patterns are reproduced in attitude-reports. First, if we replace ‘be’ with ‘be identical to’ the asymmetry of (14a)/(14b) appears to vanish:

> (14a=) Biron thinks Katherine is identical to Rosaline.
> (14b=) Biron thinks Rosaline is identical to Katherine.

At the *Masked Ball*, (14a=) lacks the true reading that (14a) has. And (14a=) and (14b=) appear equivalent.

Second, the conjunction of (14a) and the negation of (14a=), namely

> (15) Biron thinks Katherine is Rosaline, but he does not think Katherine is identical to Rosaline.

strikes me as true at the *Masked Ball*. This is a remarkable result. If ‘be’ and ‘be identical to’ both expressed numerical identity, the truth of (15) is very hard to explain.

The natural conclusion is that 'be' in equative clauses does not always denote IDENTITY, and that WH-IDENTITY and IDENTITY encode genuinely distinct relations.

My WH-IDENTITY proposal is revisionary. However, as I argue in Paper III, it is doubtful whether any non-revisionary account can explain the asymmetry phenomenon.[96] The phenomenon forces us to confront difficult theoretical choices, and I propose that the WH-IDENTITY proposal is the best option available.

Before closing, let me briefly note a historically important discussion absent from Paper III.[97] In debates about imagination, some authors discussing sentences like (5a) and (5b) have suggested that 'be' does not carry its ordinary meaning. Instead, it has been suggested that in the context of imagination reports, 'be' takes on a different sense. Vendler (1979) proposed that (5a) says that I imagine having Napoleon's experiences; Velleman (1996) that I imagine taking Napoleon's perspective; Williams (1966) that I imagine acting as Napoleon. These paraphrases seem to capture different aspects of what may be involved in imagining being someone else, and in that respect they have independent interest.

However, as seen, my concern is not with imagination itself, but with the asymmetry phenomenon – and how this linguistic pattern manifests in different constructions. From a linguistic perspective, the question is whether these paraphrases shed light on the asymmetry phenomenon. They do not, which is evident from their failure to generalize: they offer no plausible reading of belief reports such as (14a). Strictly speaking, (14a) has no interpretation on which it says that Biron thinks that Katherine has Rosaline's experiences, or that Katherine adopts Rosaline's perspective, or that Katherine is acting like Rosaline. In short, while these proposals may illuminate what it is to imagine being someone else, they do not illuminate the asymmetry phenomenon.

As I noted earlier, post-Fregean semantics has grown rich in its treatment of belief, names, and related phenomena, while the copula itself has received comparatively little sustained philosophical attention. My exploration of the asymmetry phenomenon is meant to reopen this somewhat neglected line of inquiry. Much remains to be clarified about the roles of 'be' role in natural

[96] For discussion of these alternatives, see Paper III. Briefly: the Predication hypothesis treats 'be' in sentences like (12a) and (12b) as PREDICATION. The problems are two-fold. First, since predicative 'be' is semantically vacuous, the hypothesis leaves the fact that equative clauses can be true unexplained. Second, the hypothesis lacks independent linguistic support. Next, the IDENTITY hypothesis treats 'be' as IDENTITY in sentences such as (12a) and (12b). While this works for some data, it fails to account for the full range of asymmetry data (e.g. the desire reports (4a)/(4b)).

[97] See Velleman (1996), Vendler (1979), Walton (1990), Williams (1966) and Wollheim (1984).

language. I believe that future work on the copula will not only illuminate the asymmetry phenomenon but also shed new light on the interface between language and reality.

Let us take stock. The discussions in this introduction have traced a set of problems in the epistemology and semantics of the mental – ranging from explananda and theories of introspective knowledge, to the phenomenon of IEM and identification-freedom, to the semantics of the copula. Each has been considered in relative isolation, and the papers that follow take them up in the same spirit. The phenomena are distinct and should be treated as such. Still, as this introduction has hopefully made clear, there is important thematic overlap among them. A better understanding of the asymmetry phenomenon, for instance, should inform the discussion of misidentification – clarifying how misidentifying Katherine as Rosaline differs from misidentifying Rosaline as Katherine. Here the semantics of 'be' holds particular importance. Further, since (mis)identification is thematically connected to identification-free knowledge, an important case of which is introspective first-person knowledge, the two strands of discussion are not as distant as they may initially appear. Despite the apparent heterogeneity of the topics surveyed, their interconnections provide the background for the discussions that follow.

Literature

Alston, William (1971) “Varieties of Privileged Access”, *American Philosophical Quarterly* 8:3, pp. 223—241

Armstrong, David M (1993 [1968]) *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul)

Armstrong, David M & Norman Malcolm (1984) *Consciousness & Causality: A Debate on the Nature of Mind* (New York: Basil Blackwell Publisher)

Bar-On, Dorit (2004) *Speaking My Mind: Expression and Self-Knowledge* (Oxford: OUP)

Bar-On, Dorit & Drew Johnson (2025) “’Transparent’ Rules and Basic Self-Knowledge: A Critical Study of Alex Byrne’s *Transparency and Self-knowledge*”, *Inquiry,* pp. 1—43 https://doi.org/10.1080/0020174X.2025.2476219

Bar-On, Dorit & Crispin Wright (2024) *Expression and Self-Knowledge* (Hoboken, NJ: Wiley Blackwell)

Berker, Selim (2008) “Luminosity Regained”, *Philosophers’ Imprint* 8, pp. 1—22.

Bilgrami, Akeel (2006) *Self-Knowledge and Resentment* (Cambridge, MA: Cambridge University Press)

Block, Ned (1995) “On a Confusion About a Function of Consciousness”, *Behavioral and Brain Sciences*, 18:2, pp. 227–247.

Boghossian, Paul (1989) “Content and Self-Knowledge” in his *Content and Justification: Philosophical Papers* (New York & Oxford: OUP, 2008), pp. 139—158

Boyle, Matthew (2009) “Two Kinds of Self-Knowledge”, *Philosophy and Phenomenological Research* 78:1, pp. 133—164

— (2011) “Transparent Self-Knowledge”, *Proceedings of the Aristotelian Society Supplementary Volume* 85, pp. 223—241

Butler, Jesse (2013) *Rethinking Introspection: A Pluralist Approach to the First-Person Perspective* (London: Palgrave Macmillan)

Burge, Tyler (1988) “Individualism and Self-Knowledge” in his *Cognition Through Understanding* (New York & Oxford: OUP, 2013), pp. 55—67

— (1993) “Content Preservation” in his *Cognition Through Understanding*, pp. 229—253

— (1996) “Our Entitlement to Self-Knowledge” in his *Cognition Through Understanding*, pp. 68—87

— (1998) “Reason and the First Person” in his *Cognition Through Understanding*, pp. 383—406

— (2011) “Self and Self-Understanding: The Dewey Lectures” in his *Cognition Through Understanding*, pp. 123—227

Byrne, Alex (2005) “Introspection”, *Philosophical Topics* 33:1, pp. 79—104

— (2018) *Transparency and Self-knowledge* (New York & Oxford: OUP)

Cappelen, Herman & Josh Dever (2013) *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person* (Oxford: OUP)

Carruthers, Peter (2010) “Introspection: Divided and Partly Eliminated”, *Philosophy and Phenomenological Research* 80:1, pp. 76—111
— (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (New York & Oxford: OUP)
Chisholm, Roderick (1942) “The Problem of the Speckled Hen”, *Mind* 51:204, pp. 368—373
Churchland, Paul M (2013) *Matter and Consciousness* 3rd ed. (Cambridge, MA: MIT Press)
Coliva, Annalisa (2006) “Error Through Misidentification: Some Varieties”, *Journal of Philosophy* 103:8, pp. 403—425
— (2008) “Peacocke’s Self-Knowledge”, *Ratio* 21:1, pp. 13—27
— (2016) *Varieties of Self-Knowledge* (London: Palgrave Macmillan)
— (2017) “How to be a Pluralist about Self-Knowledge” in *Epistemic Pluralism* (eds.) Annalisa Coliva & Nikolaj Jang Lee Pedersen (London: Palgrave MacMillan), pp. 253—284
— (2019) “Replies to Commentators”, *Philosophia* 74, pp. 343—352
Coliva, Annalisa & Michele Palmira (2025) “Immunity to Error Through Misidentification: Some Trends”, *Philosophical Psychology* 38:3, pp. 1126—1161
Conee, Earl (2005) “The Comforts of Home”, *Philosophy and Phenomenological Research* 70:2, pp. 444—451.
Conee, Earl & Richard Feldman (2004) *Evidentialism: Essays in Epistemology* (New York & Oxford: OUP)
Cumming, Samuel (2008) “Variabilism”, *The Philosophical Review* 117:4, pp. 525—554
Davidson, Donald (1987) “Knowing One’s Own Mind”, *Proceedings and Addresses of the American Philosophical Association* 60:3, pp. 441—458
Den Dikken, Marcel (2006) *Relators and Linkers: The Syntax of Predication, Predicate Inversion and Copulas* (Cambridge, MA: MIT Press)
Descartes, René (1985a [1637]) “Discourse on the Method” in *The Philosophical Writings of Descartes* Vol. 1. Transl. John Cottingham, Robert Stroothof and Dugald Murdoch (London & New York: Cambridge University Press), pp. 111—176
— (1985b [1641]) “Meditations of First Philosophy” in *The Philosophical Writings of Descartes* Vol. 2. Transl. John Cottingham, Robert Stroothof and Dugald Murdoch (London & New York: Cambridge University Press), pp. 12—62
Duncan, Matt (2015) “We Are Acquainted With Ourselves”, *Philosophical Studies* 172:9, pp. 2531—2549
— (2019) “The Self Shows Up in Experience”, *Review of Philosophy and Psychology* 10:2, pp. 299—318
Dretske, Fred (1995) *Naturalizing The Mind* (Cambridge, MA: MIT Press)
— (2003) “How Do You Know You are Not a Zombie?” in *Privileged Access: Philosophical Accounts of Self-Knowledge* (ed.) Brie Gertler (Ashgate: Routledge), pp. 1—14
Enoch, David & Joshua Schechter (2008) “How Are Basic Belief-Forming Methods Justified?”, *Philosophy and Phenomenological Research* 76:3, pp. 547—579
Evans, Gareth (1982) *The Varieties of Reference* (Oxford: OUP)
Farkas, Katalin (2008) *The Subject’s Point of View* (Oxford: OUP)
Fernández, Jordi (2013) *Transparent Minds: A Study of Self-Knowledge* (Oxford: OUP)
Frege, Gottlob (1948 [1892]) “On Sense and Reference”, *Philosophical Review* 57:3, pp. 209—230

Gallois, André (1996) *The World Without, The Mind Within: An Essay on First-Person Authority* (Cambridge: Cambridge University Press)
García-Carpintero, Manuel (2018) "De Se Thoughts and Immunity to Error Through Misidentification", *Synthese* 195:8, pp. 3311–3333
Gertler, Brie (2011) *Self-Knowledge* (London & New York: Routledge)
— (2012) "Renewed Acquaintance" in *Introspection and Consciousness* (eds.) Declan Smithies & Daniel Stoljar (Oxford: OUP), pp. 93—129
— (2018) "Self-Knowledge and Rational Agency: A Defense of Empiricism", *Philosophy and Phenomenological Research* 91:1, pp. 91—109
Graff Fara, Delia (Ms.) "Names as Predicates", (http://www.princeton.edu/~harman/Courses/PHI534-2012-13/Jeshion/Fara-namesaspreds.pdf. retrieved 2019-11-12)
Higgins, Francis Roger (1979) *The Pseudocleft Construction in English* (New York: Garland Publishing)
Hill, Christopher (1988) "Introspective Awareness of Sensations", *Topoi* 7, pp. 11—24
Hofweber, Thomas (2005) "Number Determiners, Numbers, and Arithmetic", *The Philosophical Review* 114:2, pp. 179—225
Horgan, Terence (2012) "Introspection about Phenomenal Consciousness: Running the Gamut from Infallibility to Impotence" in *Introspection and Consciousness* (eds.) Smithies & Stoljar, pp. 405—422
Howell, Robert J. (2023) *Self-Awareness and the Elusive Subject* (Oxford & New York: OUP)
Hume, David (2000 [1739-40]) *A Treatise of Human Nature* (eds.) David Fate Norton & Mary J. Norton (Oxford & New York: OUP)
Kang, Shao-Pu (2022) "Shared Consciousness and Asymmetry", *Synthese* 200: 413 https://doi.org/10.1007/s11229-022-03890-w
Kant, Immanuel (1998 [1781/1787]) *Critique of Pure Reason,* Translated and edited by Paul Guyer and Allen W. Wood (Cambridge: Cambridge University Press)
Keeling, Sophie (2019) "Book Review: Alex Byrne *Transparency and Self-Knowledge*", *Philosophical Quarterly* 70:280, pp. 639—642
Kocurek, Alexander (2018) "Counteridenticals", *The Philosophical Review* 127:3, pp. 323—369
Kripke, Saul (1981) *Naming and Necessity* (Malden MA: Blackwell Publishing)
— (2005) "Russell's Notion of Scope", *Mind* 114:456, pp. 1005—1037
Lewis, C.I. (1929) *Mind and The World Order: Outline of a Theory of Knowledge* (New York: Charles Scribners).
Locke, John (1975 [1690]) *An Essay Concerning Human Understanding* (ed.) P. H. Nidditch (Oxford: Clarendon Press)
Lycan, William (1996) *Consciousness and Experience* (Cambridge, MA: MIT Press)
Lyons, Jack (2019) "Algorithm and Parameters: Solving the Generality Problem for Reliabilism", *The Philosophical Review* 128:4, pp. 463—509
Manley, David (2007) "Safety, Content, Apriority, Self-Knowledge", *The Journal of Philosophy* 104:8, pp. 403—423
Macdonald, Cynthia (1999) "Shoemaker of Self-Knowledge and Inner Sense", *Philosophy and Phenomenological Research* 59:3, pp. 711—738
McHugh, Connor (2012) "Reasons and Self-Knowledge" in *The Self and Self-Knowledge* (ed.) Annalisa Coliva (Oxford & New York: OUP), pp. 139—163
Mikkelsen, Line (2011) "Copular Clauses" in *Semantics: An International Handbook of Natural Langue Meaning* (eds.) Klaus von Heusinger, Claudia Maienborn & Paul Portner (Berlin: De Gruyter Mouton), pp. 1805—1829

Moltmann, Friederike (2012) “Tropes, Bare Demonstratives, and Apparent Statements of Identity”, *Nous* 47:2, pp. 346—370
— (2013) “Identificational Sentences”, *Natural Language Semantics* 21:1, pp. 43—77
Moran, Richard (2001) *Authority and Estrangement* (Princeton NJ: Princeton University Press)
— (2003) “Responses to O’Brien and Shoemaker”, *European Journal of Philosophy* 11:3, pp. 402—419
Morgan, Daniel (2012) “Immunity to Error Through Misidentification: What Does It Tell Us About the De Se?” in *Immunity to Error Through Misidentification* (eds.) Simon Prosser & François Recanati (Cambridge: Cambridge University Press), pp. 104—123
Moro, Andrea (1997) *The Raising of Predicates: Predicative Noun Phrases and the Theory of Clause Structure* (Cambridge: Cambridge University Press)
Neta, Ram (2011) “The Nature and Reach of Privileged Access” in *Self-Knowledge* (ed.) Anthony Hatzimoysis (New York & Oxford: OUP), pp. 9—32
O’Brien, Lucy (2012) “Knowledge of Actions and Tryings” in *The Self and Self-Knowledge* (ed.) Coliva, pp. 164—179
Parrott, Matthew (2017) “Self-Blindness and Self-Knowledge”, *Philosophers’ Imprint* 17, pp. 1—22
Partee, Barbara (1986) “Ambiguous Pseudoclefts with Unambiguous Be”, in *Papers from the Parasession on Pragmatics and Grammatical Theory* (eds.) Anne Farley, Peter Farley & Karl-Erik McCullough (Chigago Linguistic Society), pp. 354—366
— (2010) “Specificational Copular Sentences in Russian and English” in *Russian in Contrast* (eds.) A. Grønn & I. Marijanovic (Oslo Studies in Language 2:1), pp. 25—49
Peacocke, Christopher (1999) *Being Known* (New York & Oxford: OUP)
— (2012) “Conscious Events and Self-Ascriptions: Comments on Heal and O’Brien” in *The Self and Self-Knowledge* ed. Coliva, pp. 180—186
Pryor, Jim (1999) “Immunity to Error through Misidentification”, *Philosophical Topics* 26:1&2, pp. 271—304
Quine, Willard van Orman (1960) *Word and Object* (Cambridge, MA: MIT Press)
Roelofs, Luke (2019) *Combining Minds: How to Think About Composite Subjectivity* (New York & Oxford: OUP)
— (2024) “No Such Thing as Too Many Minds”, *Australasian Journal of Philosophy* 102:1, pp. 131—146
Romero, Maribel (2005) “Concealed Questions and Specificational Subjects”, *Linguistics and Philosophy* 28:6, pp. 687—737
Russell, Bertrand (1912) *The Problems of Philosophy* (London & New York: OUP)
— (1919) *Introduction to Mathematical Philosophy* (London: George Allen & Unwin)
Ryle, Gilbert (1949) *The Concept of Mind* (London: Hutchinson)
Schechter, Joshua & David Enoch (2006) “Meaning and Justification: The Case of Modus Ponens”, *Nous* 40:4, pp. 687—715
Schlenker, Philippe (2003) “Clausal Equations (A Note on the Connectivity Problem)”, *Natural Language and Linguistic Theory* 21, pp. 157—214
Sharvit, Yael (1999) “Connectivity in Specificational Sentences”, *Natural Language Semantics* 7:3, pp. 299—339
Silins, Nicholas (2013) “Introspection and Inference”, *Philosophical Studies* 163:2, pp. 291—315

Shoemaker, Sydney (1968) "Self-Reference and Self-Awareness", *Journal of Philosophy,* 65:19, pp. 555-567
— (1975) "Functionalism and Qualia" in his *Identity, Cause, and Mind: Philosophical Essays* 2nd expanded ed. (Oxford: Clarendon 2003), pp. 184—207
— (1981) "Some Varieties of Functionalism" in *Identity, Cause, and Mind*, pp. 261—286
— (1996) *The First-Person Perspective and Other Essays* (Cambridge: Cambridge University Press)
— (1999) "Reply to Cynthia Macdonald", *Philosophy and Phenomenological Research* 59:3, pp. 739—745
— (2009) "Self-Intimation and Second Order Belief", *Erkenntnis* 71:1, pp. 35—51
— (2011) "Personhood and consciousness" in *Consciousness and the Self: New Essays* (eds.) JeeLoo Liu & John Perry (New York & Cambridge: Cambridge University Press), pp. 198—213
Smithies, Declan (2019) *The Epistemic Role of Consciousness* (New York & Oxford: OUP)
Stalnaker, Robert (1999) *Context and Content: Essays on Intentionality in Speech and Thought* (New York & Oxford: OUP)
— (2008) *Our Knowledge of The Internal World* (New York & Oxford: OUP)
Srinivasan, Amia (2013) "Are We Luminous", *Philosophy and Phenomenological Research* 90:2, pp. 294—319
Stoljar, Daniel (2012) "Introspective Knowledge of Negative Facts", *Philosophical Perspectives* 26:1, pp. 389—410
— (2018) "Introspection and Necessity", *Nous* 52:2, pp. 389—410
— (2019) "Evans on Transparency: A Rationalist Account", *Philosophical Studies* 176:8, pp. 2067—2085
Strawson, P.F. (1959) *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen)
— (1966) *Bounds of Sense: An Essay on Kant's Critique of Pure Reason* (London: Methuen)
Sutton, C.S. (2014) "The Supervenience Solution to the Too-Many-Thinkers Problem", *The Philosophical Quarterly* 64: 257, pp. 619—639
Velleman, David (1996) "Self to Self", *Philosophical Review*, 105:1, pp. 39—76
Vendler, Zeno (1979) "Vicarious Experiences", *Revue de Métaphysique et Morale* 84:2, pp. 161—173
Walton, Kendall (1990) *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (Cambridge, MA: Harvard University Press)
Weatherson, Brian (2004) "Luminous Margins", *Australasian Journal of Philosophy* 82:3, pp. 373—383
Williams, Bernard (1966) "Imagination and the Self" in *Studies in the Philosophy of Thought and Action* (ed.) P.F. Strawson (OUP: Oxford & New York, 1968), pp. 192—213
Williams, Edvin (1983) "Semantic Vs. Syntactic Categories"*, Linguistics and Philosophy* 6:3, pp. 423—446
Williamson, Timothy (2000) *Knowledge and Its Limits* (Oxford: OUP)
Wittgenstein, Ludwig (1958a [1953]) *Philosophical Investigations* (Oxford: Basil Blackwell)
— (1958b) *The Blue and Brown Books: Preliminary Studies for the* "*Philosophical Investigations*" (New York: Harper & Row)
Wollheim, Richard (1984) *The Thread of Life* (Cambridge, MA: Harvard University Press)

Wright, Crispin (1989) “Wittgenstein’s Rule-Following Considerations and the Central Project of Theoretical Linguistics” in *Reflections on Chomsky* (ed.) Alexander George (Oxford: Blackwell), pp. 233—264
— (2001) *Rails to Infinity: Essays on Themes from Wittgenstein’s Philosophical Investigations* (Cambridge, MA: Harvard University Press)
— (2012) “Reflections on François Recanati’s ‘Immunity to error through misidentification: what it is and where it comes from’” in *Immunity to Error Through Misidentification* (eds.) Prosser & Recanati, pp. 247—280
— (2015) “Self-Knowledge: The Reality of Privileged Access”, in *Externalism, Self-Knowledge and Skepticism* (ed.) Sanford Goldberg (Cambridge: Cambridge University Press), pp. 49—74